THE EFFECT OF VOCABULARY USE REMINDERS ON L2 ENGLISH WRITING TASKS

A Dissertation
Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Education

By
Aeric Wong
May, 2020

Examining Committee Members

Tamara Swenson, Advisory Chair, Osaka Jogakuin University
David Beglar, Teaching and Learning
Paul Nation, External Member, Victoria University of Wellington
Tomoko Nemoto, Teaching and Learning

ABSTRACT

Transfer of learning is the reason for education. The ultimate goal of teaching is to help learners apply what is studied in one situation to another. For language learning, the purpose of studying grammar and vocabulary is to be able to use that knowledge in a wide variety of contexts. It is, therefore, of interest to determine what can be done to influence transfer of learning.

In this study I explored the effects of reminders on the transfer of vocabulary learned in one course to writing tasks in another. The five purposes of this study were to evaluate the effects of the treatment on: (a) the salience of the reminders and on study time, (b) differences in vocabulary learning, (c) the developmental trajectories of language proficiency and writing based, (d) the correctness of the target vocabulary used in a writing task, and (e) the perceived usefulness of the vocabulary study.

This study was conducted at a private university in western Japan. The participants (N = 179) were enrolled in a course focusing on vocabulary learning (vocabulary course) and one on genre writing (content course). They were placed into one of four conditions based on entry-level placement tests. These groups received different degrees of reminders. Those in the comparison condition received no reminder (NR) to use the target vocabulary. The three remaining conditions received one of three

reminder treatments in the content course before revising required essays: (a) a general reminder (R) to use words studied in the vocabulary course, (b) the same general reminder and a noticing task (R+), and (c) the same general reminder and a task designed to promote target vocabulary usage (R++).

Data for this study were obtained using the following instruments: vocabulary self-study logs, Quizlet (2017) and Moodle (2016) metrics, receptive and productive vocabulary tests, a language proficiency test, three measures of writing quality, and individual interviews with 20 of the participants. Prior to conducting quantitative analyses on the data gathered with these instruments, the receptive and productive vocabulary tests, and one measure of writing quality, the JACET writing rubric (Kuru, Masaki, & Kinshi, 2009), were analyzed using the Rasch rating-scale model in order to evaluate the validity and reliability of the instruments as well as to obtain estimates of the participants' abilities and item difficulties. In addition, the Rasch model was used to check for interrater reliability and rater severity of the scores of the JACET writing rubric. The data were then analyzed using mixed ANOVAs in order to ascertain differences in within-subjects and between-subjects measures.

The results confirmed the salience of the reminders, and study time was shown to decrease in stronger conditions. This is a key finding because analysis of the data also indicated that the participants in stronger treatment conditions learned the target vocabulary with the same efficiency as those in weaker conditions despite reporting less study time. Furthermore, the treatments had positive effects on lexical complexity, amount of produced written text, and the amount of correct vocabulary usage in the writing task.

ACKNOWLEDGMENTS

First and foremost, my deepest gratitude goes to Dr. Tamara Swenson, my dissertation advisor, who has guided me since before this research project began. In fact, it was through her guidance in a research apprenticeship that the topic for this study emerged. She has been motivating, encouraging, patient throughout this process, and her keen eye for detail and invaluable advice contributed significantly to this body of work. I believe that the things she has taught me has enabled me to grow and develop both as a researcher and educator and will help me for the rest of my career. Even when things were not going smoothly, she remained a calming presence without whom I almost certainly would not have finished the dissertation.

Next, I would like to thank Dr. David Beglar, who was always available to talk to regarding my progress even though some of my visits were unscheduled. On several occasions, even going back to my time in the master's program, he provided some much-needed guidance when I was at a crossroads.

I would also like to thank the other members of my dissertation defense committee. Professor Paul Nation taught me several times through the doctoral program and his insights and extensive knowledge of our field was an impetus for this study. Dr.

Tomoko Nemoto had valuable insights and comments and significantly strengthened this study.

I am exceedingly grateful to Dr. Torrin Shimono who assisted with rating and was always available to consult regarding statistics. As someone who was also going through the dissertation-writing process at the same time, he helped me keep my eyes on the finish line. I also owe my thanks to my colleagues. Brent Jones and Roger Palmer, for allowing me to conduct my study at the university, and my peers, Ryan Richardson, Michael Hollenback, Takara Allal, Tim Ellsworth, Michael Collins, Scott Houghton, Chan Wanduragala, and David Kolf for their assistance at various stages. Also, my fellow cohort members were a constant source of inspiration.

Furthermore, I would like to acknowledge the assistance I received with the transcription and translation of the interviews. Thank you, Maki Kobayashi, for sacrificing many hours despite your busy schedule.

Finally, my heartfelt thanks to my wife, Etsuko, and my son, Micah, for their boundless patience and understanding for the many, many hours that were spent on this manuscript. They provided regular and much-needed distractions that helped me recharge, and more importantly, reminded me precisely why I was pursuing a doctorate.

Dedicated to my parents,

Dr. Wong Hong Meng & Mrs. Irene Wong

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Background of the Issue

Human beings are driven to communicate. In all societies, communication serves several purposes, ranging from the dissemination of information to the formation of social bonds. People communicate to convey facts, ideas, feelings, and even culture. Regardless of the language, communication is not possible without a vocabulary. The importance of vocabulary, therefore, cannot be overstated. An extensive vocabulary facilitates self-expression and aids understanding. Indeed, words are the building blocks of language. A conservative estimate indicates that native English speakers typically know 20,000 word families (Nation & Waring, 1997). Conversely, a small or underdeveloped vocabulary impairs precision of expression and hampers understanding. Research has shown that vocabulary is closely linked to IQ, achievement tests, and reading comprehension (Baumann, 2005). It has also been suggested that vocabulary is more important than grammar, as it is key for understanding what is heard or read and is critical to the success or failure of communication (Wilkins, 1972). Wilkins famously stated that “without grammar, very little can be conveyed, without vocabulary, nothing can be conveyed" (pp. 111-112). Similarly, Lewis (1993) argued that "lexis is the core or heart of language" (p. 89), especially throughout the entire language development process as acquisition of more productive vocabulary is central to greater fluency and expression. One only has to observe language learners to see, as Schmitt (2010) noted, that "learners carry around dictionaries and not grammar books" (p. 4). It is, therefore, of interest to explore the various avenues through which vocabulary is acquired.

Second language instructors inherently understand that vocabulary is one of the cornerstones of any curriculum, with some dedicating a substantial portion of a language course to vocabulary building. Essentially, one of the main purposes for this vocabulary component is to support students' overall language learning, which, in turn, implies transfer of knowledge to other skill areas.

For language learners, vocabulary is an ongoing struggle. Vocabulary knowledge is incremental by nature (Jones & Haywood, 2004) and some gains might be so small that they are almost impossible to ascertain. Even those who score highly on standardized tests might consider their vocabulary inadequate or, perhaps, it is more accurate to say that they feel that they have an inadequate vocabulary because they cannot use it freely (Wong, 2017). This perceived inadequacy, then, raises the question as to what words should be studied, and how best to learn them. Specifically addressing the issue of retention, Craik and Lockhart's (1972) Levels of Processing framework is an attempt at explaining why something that is processed more deeply and thoughtfully is remembered more easily. This framework was further developed by Hulstijn and Laufer (2001) to show how different tasks had different involvement loads and, therefore, could be designed to best promote vocabulary learning.

It is undisputed that some words are more valuable to study than others. Certain words occur more frequently and language learners see, hear, and need to use these high-frequency words more than they do others. However, even with high-frequency vocabulary, there are different aspects of word knowledge that can make vocabulary learning difficult as they require different approaches for effective learning (Nation, 2001). Furthermore, the use of strategies that promote long-term retention (e.g., spaced learning) is also a factor in vocabulary acquisition (Kornell, 2009). This vocabulary knowledge can be measured using a variety of receptive or productive tests that show the facility a learner has with certain vocabulary.

The next step after learning and remembering vocabulary is to use it productively through speaking and writing. Merely studying vocabulary does not automatically develop the ability to use those words. Some words might be known but, due to confusion or time pressure, learners are not motivated to use them (Corson, 1985; Laufer, 2003b; Nation, 2001). Furthermore, language is learned through experimentation. Usage based theory (Tomasello, 2000) includes the premise that language development is a function of language use and that through usage and exposure to language in a variety of contexts, grammatical patterns and structures are acquired. Therefore, language learners make better judgments regarding the appropriate use of vocabulary the more they use this vocabulary and, simultaneously, obtain feedback on its use (Tomasello, 2000). With the understanding that using and experimenting with language is beneficial for language development, language learners also need to notice when they fail to communicate successfully. Essentially, they need to be pushed to develop communication strategies that help them find alternative ways of expressing meaning when their productive vocabulary is insufficient. By doing so, they might learn to use words they know in creative ways, that is, not in the way they originally learned those words (Nation & Webb, 2011). Using words when speaking and writing has the potential benefit of helping learners turn their receptive vocabulary into productive vocabulary. As they become more confident through using the vocabulary, they transition from using a word because they have to in a specific task or activity to being able to use a word voluntarily.

Ultimately, the goal of any learning is the transfer of knowledge. In the case of language learning, transfer of knowledge can take place across time, meaning that successful learning is demonstrated by applying knowledge in future situations, similar or new. For vocabulary learning, transfer of knowledge can be from receptive to productive use, as in, from only understanding words when listening or reading to being able to use the words through oral or written communication. In the same way, transfer of vocabulary knowledge can be from explicit study of words to use of those previously studied in writing. Therefore, it is important to understand how words are retained in memory, what affects long-term retention, and the processes involved in transfer of vocabulary studied to use of this vocabulary.

Statement of the Problem

I address five problems in this study. The first problem concerns the lack of second language (L2) research regarding the general salience of treatments as a contributing factor on the effects obtained. I was unable to locate any studies of that investigated the salience of different treatments to L2 learners. This is an important consideration as a failure to notice or understand the treatment would render any findings regarding the effects of the treatment inconclusive. This is especially relevant for studies that employ the use of more subtle treatments such as simple reminders and the effects of these on vocabulary study time. Hence, more research is needed concerning the salience of treatments in experimental studies.

The second problem is that SLA research has not clearly shown the contribution of study time combined with study methods to productive vocabulary use. A large body of research exists that describes the effects of study frequency (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1987; Kornell, 2009), and study methods such as using word lists on vocabulary learning(e.g., Nakata, 2011). However, I did not locate any studies which have examined the effects of a combination of study time and one or more study methods on vocabulary learning and use. Therefore, it is clear that studies are needed to investigate the interaction between study time and study methods in terms of their efficacy in promoting vocabulary learning and use.

The third problem is that while many studies have shown a strong relationship between vocabulary knowledge and other L2 skills (e.g., Hilton, 2008, for speaking; Jordan, 1997, for writing; Laufer, 1997, for reading comprehension; Miralpeix & Muñoz, 2018, for reading, writing, listening, and speaking), researchers have not examined extensive causal sequences, such as from study time to vocabulary learning to language proficiency or writing quality. In other words, there is still a lack of clarity regarding the roles of various factors on the constructs of knowledge, learning, and language proficiency.

The fourth problem is that few researchers have presented findings regarding the degree of correct use of vocabulary and how this contributes to writing quality. Several researchers have established that vocabulary is a good predictor of writing scores (e.g., Astika, 1993; Jordan, 1997; Santos, 1988). Others have investigated different features of writing (e.g., Crossley, Kyle, & McNamara, 2016, cohesive devices; Crossley & McNamara, 2014, syntactic complexity; Laufer, 1994, lexical richness; Ong & Zhang, 2010, lexical complexity). However, these studies have not included assessments of whether the target vocabulary was used correctly, partially correctly, or incorrectly. Consequently, findings regarding vocabulary knowledge and the degree to which words are used correctly remain inconclusive.

The fifth problem is that few researchers have addressed how perceived usefulness affects the findings in experimental treatments. This problem is concerned with the intended purpose of vocabulary study. As tests are the de facto method to measure progress, it is easy for language learners to forget that vocabulary study is intended to support overall language development. There is a wide body of research describing motivation and its effects on learning; however, there is a lack of research concerning how perceived usefulness affects the interpretations of the results.

Purposes and Significance of the Study

The overall purpose of this study was to investigate the effect that reminders have on learning and whether this resulted in vocabulary transfer between courses. This study used longitudinal quantitative data collection methods. There were five specific purposes.

First, the study investigated whether the students noticed the reminders. Establishing salience was critical as any subsequent findings regarding study time or the developmental trajectories of vocabulary knowledge, language proficiency, or writing quality would be affected. This study incorporated three reminder treatments over one academic semester: a vocabulary use reminder given orally and in writing, the reminder plus a task which encouraged the participants to choose potentially useful words for their essays, and the reminder plus a task which required the participants to write the chosen vocabulary in a provided space above the final drafts of their essays. A control group with no reminders was also included in the study. It was, therefore, of interest to evaluate the relative effectiveness of these three reminder treatments. The first purpose is significant to SLA, researchers and adds to the body of scholarly literature in the field because results regarding the factors which influence the rate of learning remain inconclusive. As a result, more studies are necessary to substantiate past findings. The results of this study might also allow educators and curriculum designers to better understand how to improve their students' vocabulary learning and use. With a more complete understanding of the effect of reminders on language learning, in general, and the learning of vocabulary, educators can incorporate reminders at optimal times. The results of the study might also influence decisions regarding cooperation between teachers in coordinated programs where separate courses are designed with interrelated goals and learning objectives.

The second purpose of this study was to assess the degree to which study time and study methods contributed to vocabulary learning and retention. For example, it was necessary to compare the amount of reported study time with gains in vocabulary test scores. Similarly, comparisons were also drawn between different types of study methods and vocabulary test scores. Examining the relationship between study time, study methods, and vocabulary learning was important as it contributed to interpreting the results regarding the productive use of the vocabulary studied. This issue is significant because, despite a large body of research investigating vocabulary learning and use (e.g., Laufer & Paribakht, 1998; Hilton, 2008; Uchihara & Saito, 2019), it is still unclear if studying vocabulary automatically leads to the ability to use those words productively. This study can also add to the body of research regarding the combination of study time and study method on vocabulary learning. By having a better understanding of how much time learners dedicate to vocabulary study and the methods they employ, teachers can improve practice as they can gain insight into the amount of time it takes to develop vocabulary knowledge and the methods that prove to be efficient.

The third purpose was to investigate the connection between vocabulary knowledge, L2 language proficiency, and writing quality. However, to the best of my knowledge, no L2 researchers have examined the causal sequence from study time to vocabulary learning to language proficiency. Furthermore, as stated above, the relationship between receptive vocabulary knowledge and productive use is still not fully understood. Therefore, it was of interest to evaluate the effect that the treatments had on target vocabulary use in a writing task. The significance of the third purpose lies in its contribution to the literature of vocabulary knowledge and in attempting to establish causal relationships between study time, vocabulary learning, language proficiency, and writing quality. In addition, understanding this connection can help researchers and educators make more informed decisions regarding the type of treatments to employ to help learners develop and apply their lexical knowledge.

The fourth purpose is to assess the how vocabulary knowledge affects the degree to which words are used correctly. This purpose is significant because it can contribute to determining the link between vocabulary knowledge and ability to use words correctly. I have not located any studies examining the degree of correct usage of target vocabulary, and the results from this study can help researchers better understand how depth of vocabulary knowledge affects productive vocabulary use.

The fifth purpose was to acquire a better understanding of the perceived usefulness of L2 vocabulary study by learners in relationship to their writing. Many language programs include lists of specific words that learners are required to learn. However, the purpose of such vocabulary learning is not always clear to learners. For example, they might be directed to learn a specific list of vocabulary with the ultimate goal of enabling them to use those words accurately and appropriately in various contexts. However, if this goal is not clear, or if the students do not understand the potential benefit of the vocabulary study, successful learning is less likely to occur. Therefore, perceived usefulness has clear motivational implications which can affect learning. This issue is significant because it can add to the to the body of research regarding motivation and L2 learning beliefs. By understanding how learners perceive different tasks, researchers and educators can improve practice as they can gain insight into how best to present these tasks.

The Audience for the Study

This study will potentially benefit four audiences: second language acquisition (SLA) researchers, language teachers, curriculum designers and materials developers, and educators and administrators outside the field of SLA. The first audience for this study is vocabulary in second language acquisition (SLA) researchers. This study provides these researchers with information about the effect of increased attention, even through the use of minimal reminders, on vocabulary use. Providing this information aids researchers interested in examining whether reminders influence learners' performance and are applicable to language production in other second language learning contexts.

The second audience for this study is language teachers. As many educational institutions have a vocabulary-learning component in their language programs, it is useful to have information about the potential of focused reminders for language development. Particularly with regard to language production, language teachers should consider the long-term implications and desired outcomes of vocabulary teaching in terms of transfer of knowledge across courses. Furthermore, the task of reminding students might make this study of use to those teaching in other disciplines or age groups.

Similar to language teachers, curriculum developers and materials designers wanting to create coordinated syllabi can incorporate strategies based on the findings of this study. Curriculum developers, particularly those overseeing several courses, will find it valuable to integrate reminders into syllabi at optimal times.

Outside the field of SLA, educators and administrators, especially those who design and make curriculum decisions, should be interested in determining if even slight attention-raising produces measurable results regarding transfer of skills. This information would prove useful in considering the workload and balance of various courses that students take, and, therefore, facilitate more effective use of time.

Delimitations

There were four main delimitations in this study. First, the English proficiency of the participants limited the generalizability to learners at a similar proficiency. Results are not generalizable to English language learners at lower or higher proficiency levels because the chosen intervention, data collection instruments, and target vocabulary would need to be different. Second, the first language (L1) of the participants was Japanese. Therefore, the results of this study are not generalizable to learners with other first languages as different languages have unique patterns and rules. Third, the participants attended a single private university in western Japan. The results of this study are not generalizable to learners at universities that are ranked either higher or lower (see Chapter 2 for a discussion of the Japanese education and university rankings) as those students would have different learning histories and a different range of language ability. Finally, I only examined one type of writing (i.e., recount genre) under test conditions. These parameters, therefore, limit the applicability to other genres of writing or to writing not undertaken in test conditions.

Organization of the Study

In this study, I examined the extent of vocabulary transfer as a result of degrees of in-class reminders. The vocabulary of interest was the second 500 words of the New General Service List (NGSL) (Browne, Culligan, & Phillips, 2013b) and their occurrence in writing tasks performed under test conditions. In Chapter 2, Review of the Literature, I provide a review of the literature addressing learning and memory, the role of output in language learning, second language vocabulary learning, the components of good writing, as well as a brief overview of the Japanese educational context, followed by two pilot studies that informed the methodological decisions described in the main study. At the end of the chapter, the gaps in the literature, purposes of the study, and hypotheses are presented. In Chapter 3, Methods, I outline the context for the study, participants, instruments, research design, data analysis, and Rasch analysis. In Chapter 4, Preliminary Analyses, I describe the preliminary analyses that are necessary prior to analyzing the data for the hypotheses. In Chapter 5, Results, I present the results and their relationship to previous studies. In Chapter 6, Discussion, I provide a discussion of the findings including a summary and interpretation of the results. At the end of this chapter, I also discuss the pedagogical implications of the study. In Chapter 7, Conclusion, I summarize the main findings of the study, discuss the limitations of the present study, make suggestions for future research, and make concluding comments.

CHAPTER 2

REVIEW OF THE LITERATURE

In this chapter, I provide an overview of learning and memory, the role of output in language learning, second language vocabulary learning, and components of good writing. In these sections, I review the literature regarding some key theories, concepts, and some key empirical studies of memory and second language acquisition. Additionally, I discuss the Japanese educational context, which is useful in understanding some of the approaches or methodological decisions in this study.

Learning and Memory

Few would disagree that memory plays a crucial role in learning. However, the specific processes that facilitate or promote learning are still not fully understood. In this section, I review key theories regarding the role of memory in language learning.

Levels of Processing

One of the main issues in learning and memory revolve around both short- and long-term memory. The Levels of Processing (LOP) framework (Craik & Lockhart,

1972) focused on the retention of language. Craik and Lockhart aimed to build models of short-term and long-term memory and stated that the encoding of language was influenced by visual, phonemic, and semantic features. They argued that different encoding led to different degrees of retention. The fundamental premise of LOP was that the more deeply and thoughtfully individuals process information, the more likely it is to be retained in memory. For Craik and Lockhart, processing is either described as Type 1, which is processing at a structural or sensory level, known also as shallow processing, or Type 2, which is processing at a semantic level, or deep processing.

Within L2 education, some commonly used classroom activities such as *listen and repeat* or reading aloud with a focus on correct pronunciation fall under the sensory category, Type 1 in Craik and Lockhart's (1972) LOP framework. However, an activity such as the key word technique, a mnemonic, would be categorized as semantic level processing, Type 2, as it focuses on meaning. For example, the word *affluent* sounds like 溢れる (*afureru* = to overflow). Therefore, the word *affluent* can be linked to a person in a bathtub overflowing with money. In LOP theory, Type 2 processing leads to a more durable memory trace and, in short, better retention.

The hierarchical stages of the LOP framework (Craik & Lockhart, 1972) means that a greater depth of processing implies a greater degree of semantic analysis and that memory traces are a function of depth of analysis. This, in turn, suggests that people need to remember the meanings of stimuli (i.e., words) whereas preliminary superficial analysis (e.g., on the spelling or pronunciation) is less important. As such, retention is a function of depth or degree of elaboration, amount of attention, processing time, and compatibility with prior knowledge.

A number of studies in incidental learning have demonstrated support for LOP and indicated that greater depth of processing leads to greater learning. Craik and Tulving (1975), in an effort to explore the LOP framework (Craik & Lockhart, 1972), subjected 20 participants to various experiments designed to simulate degrees of semantic involvement. Their findings led them to conclude that “retention depends critically on the qualitative nature of the encoding operations performed; a minimal semantic analysis is more beneficial for memory than an elaborate structural analysis” (p. 268). Other researchers following this line of questioning similarly concluded that:

- Students “who encoded semantically, either by producing associates or synonyms, demonstrated less distraction and more correct recognition than acoustic encoders or intentional learners.” (Elias & Perfetti, 1973, p. 155)
- “semantic tasks are conducive to high levels of recall and high degrees of organization. . . Tasks which require processing of the graphic or acoustic form of

the stimulus words result in poor recall and little associative structure" (Hyde & Jenkins, 1973, p. 478)

- "the difference between performance on the two retrieval tests was greater following semantic than rhyme encoding" (Marmurek, 1995, p. 188)

Furthermore, these studies have shown that processing at the structural level might lead to short-term but not long-term retention, while processing at the semantic level, or elaborative processing, leads to more durable, long-term retention.

Despite an extensive body of research, it should be noted that there are several criticisms of the LOP framework. First, what might be considered deep processing for one person might not be deep for another (Ekuni, Vaz, & Bueno, 2011). The nature or difficulty of the task translates to time taken thereby confounding type of processing, amount of effort, and length of time spent on processing which makes it difficult to identify and isolate any variation on the same level of processing that leads to different effects on memory (Ekuni et al., 2011). Additionally, time spent on processing has been shown to be reducible with training. Neurological scans have shown increased brain activity (i.e., effort) which correlates with higher levels of processing (e.g., Fletcher, Shallice, Frith, Frackowiak, & Dolan, 1998; Friederici, Opitz, & Von Cramon, 2000; Kapur, Craik, Tulving, Wilson, Houle, & Brown, 1994). A second criticism of

LOP points to the difficulty in teasing apart the LOP model as there is no clear delineation in the levels (Baddeley, 1978; Eysenck, 1978; Nelson, 1977). The terms deep and shallow, when referring to depth of processing, remain vague and ill-defined which makes them difficult to measure (Baddeley, 1978; Eysenck, 1978; Nelson, 1977). Baddeley (1978) and Eysenck (1978) have expressed doubt regarding a hierarchy of levels ranging from shallowest to deepest due to a lack of supporting research suggesting that there is not a fixed series of processing stages. Furthermore, there is yet to be an independent measure of depth of processing, which limits the extent to which processing can be considered shallow or deep (Lockhart & Craik, 1978). Finally, critics of the LOP framework (Eysenck, 1978; Nelson, 1977) have pointed out that the tests involved do not evaluate the theory directly but the secondary effects of those theories. That is, as a result of the vagueness with which depth is defined, retention tests are commonly used to measure the LOP (Challis, 1993; Craik & Tulving, 1975; Elias & Perfetti, 1973; Folse, 2006; Hulstijn & Laufer, 2001; Hyde & Jenkins, 1973; Marmurek, 1995), but the results obtained could be due to amount of effort, time, or type of processing not depth.

Since its inception in 1972, the original LOP framework has undergone changes and evolution to address criticisms levelled at it (Ekuni et al., 2011). Morris, Bransford, and Franks (1977) argued that semantic (deeper) processing is better than phonological (shallow) only when the retrieval task requires the remembrance of meaning (i.e., the recall task matches the guided task). Morris et al. (1977) proposed transfer-appropriate processing (TAP) as an alternative to LOP due to this perceived incompatibility. The theory of TAP posits that memory performance is better when the cognitive processes engaged during retrieval match the cognitive processes that were engaged when the material was encoded. The belief that test performance should be good if the study conditions match the test conditions is based on TAP. In response to this, Tulving (1979) developed the encoding specificity principle, which stated that TAP is not incompatible with LOP and that the LOP effect still applies despite evidence that similar encoding-recall conditions produce the best results (Morris et al., 1977). In short, Tulving (1979) argued that semantic (deeper) encoding resulted in better long-term retention than more shallow types of encoding while acknowledging encoding-recall condition compatibility. TAP was, therefore, added to the LOP framework, and the term *robust encoding* (Lockhart & Craik, 1990) was introduced to explain that "a more deeply encoded trait becomes accessible to more cues at the time of recall" (Ekuni et al., 2011, p. 335). Furthermore, the relation between LOP and neuroimaging and positron emission tomography (PET) scans (Nyberg, 2002) lent additional support to the LOP framework. In these studies, the researchers have identified LOP as an important part of learners' neurological activity that leads to retention of language.

Although the criticisms regarding the operationalization of processing depth are warranted, there is value in the LOP framework, as it raises awareness that different ways to teach might be effective to different degrees and provides an explanation as to why some information might be either more or less memorable (e.g., Ekuni et al., 2011; Hulstijn & Laufer, 2001). The contribution of the LOP framework to the understanding of human memory is substantial, particularly the findings that identify the act of remembering as a process and that the processing used at the encoding stage is crucial to remembering (Tulving, 2001).

Involvement Load Hypothesis. The Involvement Load Hypothesis (ILH) (Hulstijn & Laufer, 2001) is based on the LOP as first proposed by Craik and Lockhart (1972) and then elaborated on by Craik and Tulving (1975). The ILH addresses the criticisms that distinct levels of processing are ambiguous and not operationalisable, and that superiority of one learning condition over another due to deeper processing is not falsifiable (Nelson, 1977, p. 160). This means that if depth of processing cannot be clearly defined with clear, separate levels, then, by extension, studies cannot conclude that one learning condition is more effective than another based on LOP. The ILH, as a response to a lack of clear operationalization of levels of depth (Laufer & Hulstijn, 2001), provides a way to operationalize factors associated with deeper incidental vocabulary learning.

The ILH consists of three factors connected to vocabulary learning: need, search, and evaluation (Laufer & Hulstijn, 2001). Need refers to a motivational, non-cognitive dimension. Search and evaluation are cognitive dimensions, with the search factor referring to noticing and the evaluation factor referring to deliberately attending to vocabulary. A scenario in which the three factors are present would be determining if the unknown word is needed to complete a task, retrieving or finding new information about the word, and making choices between available options or generating a suitable context for the word. Laufer and Hulstijn (2001) used tasks designed to vary the degree of need, search, and evaluation and operationalized them into a motivational-cognitive construct of involvement. These tasks included reading and comprehension questions, as well as writing tasks (see Laufer & Hulstijn, 2001, p. 18, for a more detailed description of these tasks). After considering the different task-induced involvement loads of each task, the researchers concluded that their proposed construct supported the hypothesis that words processed with higher involvement load are better retained than words processed with lower involvement load.

While the ILH is a valuable move from theory to practice and is an initial attempt at operationalizing factors that contribute to learning, several studies have indicated that further research is needed to more fully understand the relationship between involvement load and time on task. Folse (2006), in a study of how different types of writing exercises affect vocabulary retention, had his participants practice target vocabulary in three written conditions of increasing involvements: one fill-in-the-blank exercise, three fill-in-the-blank exercises, and one original-sentence-writing exercise. His analysis revealed that the words practiced under the three fill-in-the-blank exercises were retained better than those in the other two conditions despite the original-sentence-writing exercise requiring more involvement. This finding is counter to the ILH as it suggests that the number of word retrievals is more effective than depth of processing. Similarly investigating the ILH, Keating (2008) conducted a study he varied the involvement of three vocabulary learning tasks: reading comprehension with marginal glosses, reading comprehension with blanks in place of the target words, and writing original sentences with the target words. While Keating found that his results supported the ILH, adding time on task as a factor diminished the effect of manipulating the involvement load. In light of these studies, it is suggested that more research is needed to compare tasks with the same involvement loads and different completion times, tasks with different involvement loads with the same completion time, the effect of using different parts of speech, the use of multiword units or idioms, and the combinations of different writing exercise types beyond fill-in-the-blank and writing original sentences. Moreover, various tasks have different involvement loads, which explains why they either promote or fail to promote vocabulary acquisition. Laufer and Hulstijn (2001) suggested the use of scoring criteria for need, search, and evaluation to estimate the task-induced involvement of activities.

Technique Feature Analysis. Despite the ease of quantification that the ILH allows in assessing learning conditions, Nation and Webb (2011) noted that it does not include key vocabulary-focused features that other studies have shown to be important. To address this, they proposed a different framework to operationalize depth of processing. This framework provides a checklist of criteria to examine the effectiveness of vocabulary teaching techniques. Table 1 lists the criteria for Technique Feature Analysis (TFA) as displayed by Nation and Webb (p. 7). (See Chapter 3, Methods, for the TFA scores for the instruments in this study.)

TFA is, therefore, a more detailed framework that operationalizes the construct of elaborate processing for vocabulary learning. It is different from the ILH in that it conceptualizes elaborate learning as consisting of several attentional components. Empirical studies (e.g., Hu & Nassaji, 2016) have demonstrated that TFA was able to better explain vocabulary learning gains compared to the ILH. In a more recent study by Zou and Xie (2018), the researchers developed and examined the effectiveness of a personalized vocabulary learning system based on the TFA checklist that recommends word learning tasks based on user models and TFA-based task scores. The researchers distinguish between personalized and non-personalized by defining the former as being guided by the TFA checklist (see Table 1). They found that a personalized learning system using the full checklist resulted in the best learning performance compared to a non-personalized approach, and a personalized approach guided by a partial version of the TFA. These findings further support the usage of the TFA to evaluate the effectiveness of different tasks for vocabulary learning.

Table 1. *Checklist of Criteria for Technique Feature Analysis*

Criterion	Statement	Scores	
Motivation	Is there a clear vocabulary learning goal?	0	1
	Does the activity motivate learning?	0	1
	Do the learners select the words?	0	1
Noticing	Does the activity focus attention on the target words?	0	1
	Does the activity raise awareness of new vocabulary learning?	0	1
	Does the activity involve negotiation?	0	1
Retrieval	Does the activity involve retrieval of the word?	0	1
	Is it productive retrieval?	0	1
	Is it recall?	0	1
	Are there multiple retrievals of each word?	0	1
	Is there spacing between retrievals?	0	1
Generation	Does the activity involve generative use?	0	1
	Is it productive?	0	1
	Is there a marked change that involves the use of other words?	0	1
Retention	Does the activity ensure successful linking of form and meaning?	0	1
	Does the activity involve instantiation?	0	1
	Does the activity involve imaging?	0	1
	Does the activity avoid interference?	0	1
Maximum score		0	18

Note. From *Researching and Analyzing Vocabulary* (1st ed.), by I. S. P. Nation and S. Webb, 2011, p. 7, Boston, MA: Heinle. Publsher's statement: Nation/Webb. Researching and Analyzing Vocabulary, 1E. © 2011 Heinle/ELT, a part of Cengage, Inc. Reproduced by permission. www.cengage.com/permissions.

Crammed vs. Massed vs. Spaced Learning

In the context of L2 vocabulary acquisition, a common study technique is the use of word lists or word cards (see Second Language Vocabulary Learning later in this chapter). For word cards, learners can adopt one or a combination of the following approaches: crammed, massed, or spaced. Kornell (2009) described crammed, massed, and spaced learning in terms of a large stack of flashcards: crammed learning concentrates study to just before it is needed such as just before a test, massed learning is dividing a stack of cards into smaller decks and studying each deck in separate study session, and spaced learning is studying one large stack of flashcards in a study session. Kornell investigated the differences between a massed learning group and a spaced learning group. His massed learning group was given a set of 20 cards that were divided into four equal decks of five cards each and studied a different deck in each study session. His spaced learning group studied the whole 20-card set in each study session. Kornell (2009) found that 90% of his participants learned more in the spaced condition than the massed condition while only 6% showed the reverse pattern.

In an earlier study, Dempster (1987) manipulated the amount of context given in addition to investigating the spacing effect. In a series of five experiments, his native English-speaking participants were given 38 uncommon English words presented in the form of booklets with varying amounts of context. The amount of time to study each word was controlled via prerecorded message. While the results provided no evidence that the amount of context was beneficial to vocabulary learning, they indicated that vocabulary learning was more successful in the spaced learning groups than the massed learning groups by approximately 0.4 to 1.2 standard deviations.

Despite evidence that cramming is an inefficient learning strategy (e.g., Dempster, 1987; Kornell, 2009), students continue to employ it as a way to learn or remember material that they anticipate will be on exams. From the perspective of a student whose short-term goal is to obtain a satisfactory grade, cramming might seem perfectly sufficient. However, crammed learning rarely leads to long-term retention.

Beyond the area of vocabulary learning, various researchers have looked at massed learning compared to spaced learning (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1996; Glenberg, 1979; Hintzman, 1974). Kornell (2009) applied research on spaced learning from psychology to education, specifically, on flashcard use. His reason for using flashcards was based on the assumption that they are among the most common resources used to study facts (Kornell & Bjork, 2008). In using flashcards, learners have to make decisions regarding the size of the deck of flashcards and how often to study them. A larger deck of cards would result in a longer time between encounters with each individual item, which is known as within-session spacing. The time between study sessions of that deck is referred to as between-session spacing. In using flashcards, learners often divide a deck into smaller decks, which

Kornell termed *massed learning*. This is different to the typical definition of massed learning where all learning is condensed into one session. Kornell found that keeping flashcards in one large deck was more effective as there was increased spacing (i.e., intervals) between learning opportunities of a given item. This spacing was also shown to be more effective than cramming (studying just before a test). These findings were in line with studies by Dempster (1987) and Pimsleur (1967), which showed that increasingly spaced retrievals had the best results for long-term retention. More recent research (Nakata & Webb, 2016) compared whole learning of vocabulary (studying a large block of words, or spaced learning) to part learning (the same amount but divided into smaller blocks, or massed learning). Nakata and Webb also found that spaced learning was more effective for retention. Additionally, whole learning (spaced) or part learning (massed) had little effect as long as the spacing was equivalent. Initially, this might seem at odds with Kornell's (2009) description of spacing (i.e., studying one large stack of flashcards). However, the crucial aspect here that results in better retention is not the singular act of studying one large stack (or whole learning) but the repeated use of that stack (Nakata & Webb, 2016).

The spacing effect has long been established as one of the "most robust phenomena in experimental psychology" (N. Ellis, 1995), and a number of studies investigating its effectiveness found support for it (e.g., Baddeley, 1997; Hulstijn, 2001; Mizuno, 2003). Essentially, crammed learning is less effective than massed learning, and massed learning is less effective than spaced learning. As such, if the goal of second language learning is long-term retention, the findings from existing research have established the importance and value of spaced learning strategies.

Further investigating the spacing effect, Nakata (2015) compared the effects of equal and expanding spacing on L2 vocabulary. 128 Japanese college students studied 20 English-Japanese word pairs and both types of spacing, equal or expanding, and amount of spacing were manipulated. In equal spacing, the intervals between encounters were constant. In expanding spacing, the intervals between encounters were gradually increased. For amount of spacing, there were also groups for massed, short, medium, or long spacing intervals. The results showed that expanding spacing was more effective than equal spacing. While other studies have established the superiority of expanding spacing over equal spacing (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008; Storm, Bjork, & Storm, 2010), Nakata's is the first L2 study to show similar findings.

In continuing this line of inquiry, Nakata and Suzuki (2019) evaluated the effects of massed and spaced learning on semantically related and unrelated words. L2 learners typically encounter words that are semantically related to each other. Examples include coordinates (e.g., *apple*, *orange*), synonyms (e.g., *fast*, *rapid*), or antonyms (e.g., *increase*, *decrease*). This semantic clustering has thought to facilitate learning as it reflects how vocabulary is stored in the mental lexicon (Nation, 2000) and how they are interconnected via word associations (Meara, 2009). However, other research has shown semantic clustering to be detrimental to learning due to interference (Baddeley, 1997; Schmitt, 2007). These inconsistent results are evident in other studies exploring impact of learning semantically related words simultaneously (e.g., Folse, 2004; Hoshino, 2010; Ishii, 2015). Nakata and Suzuki found that, while there were no significant differences in posttest scores between participants studying semantically related or unrelated words, semantically related items led to more interference errors than unrelated items. In addition, spaced learning benefitted the learning of unrelated items.

The next logical step after successful learning is the ability to use knowledge quickly, accurately, spontaneously, and effortlessly. DeKeyser (2007) termed this automatic knowledge. In L2, learners can learn grammatical rules and then engage in deliberate practice to develop knowledge about how to use those rules, that is, from declarative knowledge to procedural knowledge. With further practice, automatization can occur. Within the field of cognitive psychology, distributed practice has been widely accepted to be superior to massed practice. This is different to massed learning (see above) in that massed practice is described as short time intervals between training sessions (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). Distributed practice, conversely, is long time intervals between training sessions. Suzuki and DeKeyser (2015) examined the effect of distributed practice on proceduralization of an element of Japanese morphology. Findings suggest that massed practice, short time intervals between study sessions, is potentially more effective than distributed practice for proceduralization of vocabulary and grammar.

In summary, previous research has shown that, for effective, long-term second language vocabulary learning to take place, a combination of factors needs to be considered: spaced learning, expanding spacing, semantic clustering, and the time intervals between study sessions. Nakata and Suzuki (2019) found that the advantage of spaced learning for semantically unrelated words was more pronounced. Suzuki and DeKeyser's (2015) suggest that more frequent study sessions are more effective than those with longer intervals between them.

Usage-Based Theory

The premise behind usage-based theory is that language development is a function of language used in communication. Though sometimes confused with Chomsky's work on universals in language, usage-based theory is more empirically-based in which the use of the language is what shapes the structure of the language (Bybee, 2009). Nick Ellis, alone and with others, has argued for more research that will help build a dense corpora of learner audio, video, and transcripts to allow researchers to explore the actual use of language in a variety of contexts (N. Ellis, 2015).

Some of this work has been done. Tomasello (2000) explored how data-based models explain children's L1 language development and posited that linguistic skills are a function of accumulated experience with language leading to the building of an inventory of linguistic units. Tomasello argued that utterances are the item-based units people use and that such utterances are linguistic acts that are expressed from one person to another with communicative intention. Such utterances are used to express coherent and whole intentions. While it is possible to identify smaller units of language, Tomasello argued that these only play certain roles within utterances. Utterances can be fixed (e.g., *Could you please. How are you? I dunno*. *Sorry, I am* *late*.), and stored and retrieved as single units, or they can be abstract linguistic constructions that involve filling and addition (e.g., What's C doing Y?).

When their L1 is still in the early stages of development, children attempt to produce the complete utterances that they have heard (Tomasello, 2000). These attempts are often unsuccessful and result in only production of part of an utterance (holophrases). For example, children might attempt to reproduce *Where's mommy?* but will only be able to produce *mommy?* Despite being incomplete, holophrases suggest that imitation occurs. As such, Tomasello observed that language produced is a function of encounters with utterances in that it is a form of mimicry that is refined through increased exposure. Yet, it must be noted that novel utterances that reflect similar errors do not occur because they were never encountered (e.g., **Mary hit I* or **Jim kissed she*).

As they mature, children need to move from holophrases and become more syntactically competent language users (i.e., they need to move from part to whole utterances) (Tomasello, 2000). Doing so depends on the comprehension of adult speech as well as understanding the functional role of different elements in the language. Gaining this understanding is the beginning of learning grammar and this begins with words or phrases structuring utterances through filling variable slots. This type of early phrasal development is called *the pivot look* (e.g., Where's the X?, I wanna X, More X, I like X, I hate X). These utterances have some grammatical structure and include constant linguistic material that has some complexity and at least one slot that can be filled with words or phrases. Based on the feedback received from their utterances, certain expressions are deemed appropriate or inappropriate, with appropriate expressions naturally becoming used more. This increased use leads to entrenchment, which suggests that the mental push of reminders can aid entrenchment especially when coupled with high quality feedback.

Having seen how usage-based theory can account for L1 development, Eskildsen (2008) conducted a longitudinal case study of a Mexican-Spanish ESL learner. He concluded that formulas, what he also calls holophrases, or multi-word expressions depend on both the context of the interaction and the setting in which the interaction takes place. This means certain expressions are fleeting in nature and that their usage is determined by specific usage events. As such, for such formulas to be entrenched, the events in which the formulas occur must be recurrent, suggesting that there a strong social aspect also plays a role.

Based on the premise that holophrases are extracted from a speaker's experiences, then it follows that identified patterns are subject to change every time a new utterance is encountered (e.g., N. Ellis, 2002, 2012, 2015; N. Ellis, O'Donnell, & Römer, 2015; N. Ellis & Wulff, 2015; Gries & N. Ellis, 2015). This would indicate that linguistic knowledge is not fixed and is instead flexible and in a constant state of flux (Hopper, 1998). This would suggest that the bulk of language acquisition occurs due to implicit learning from usage (N. Ellis, 2002, 2012). N. Ellis (2012), in a review of studies of formulaic language acquisition, argued that "syntactic development consists of acquiring a large repertoire of constructions and formulas" and, based on experience, using the language (p. 30). Specifically, for L2 acquisition, language learning is largely based on the amount of linguistic input that learners receive. Crucially, the quantity of input allows learners to induce the rules of the L2 by employing cognitive mechanisms that work in any kind of learning.

The teaching implications of usage-based theory are twofold. One, frequency is likely to be a factor in acquiring target patterns. Therefore, quantity and variation of both meaning-focused input and output are essential. Two, deliberate instruction focusing on individual forms or sequences can lead to greater learning of those items. As such, teachers should try to use important or troublesome sequences as often as possible (e.g., Bybee, 2009; N. Ellis, 2015; Wulff & Ellis, 2008).

Frequency effects in language acquisition. Drawing upon psycholinguistic and cognitive theories of language acquisition, N. Ellis (2002) suggests that all linguistic units are a function of language use (e.g., Brent, 1999; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990; Jusczyk, 1997; Treiman & Danis, 1988). Usage-based theory suggests that grammar is the learning of thousands of sequences and the frequency-biased abstraction or regularities within them. It follows, then, that vocabulary acquisition similarly occurs through repeated encounters with words in a multitude of settings. Overall, language learning is the associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is, thus, a key determinant of acquisition.

Numerous researchers have investigated the effects of frequency across all levels of language representation: letter frequency in orthographic recall (Miller, Bruner & Postman, 1954), bigram frequencies in word identification (Broadbent, 1967), phonotactic knowledge in speech segmentation (Aslin, 1993), spelling-to-sound correspondences in reading (Coltheart, 1981), cohort effects in spoken word recognition (Marslen-Wilson, 1987), the influence of type frequency on pattern productivity (Bybee, 2003), constraint satisfaction in sentence processing (Seidenberg, 1997), and data-oriented parsing (Bod, 1992). The findings from these studies indicate that frequency affects the perception of letters, bigrams, and trigrams, as well as influencing processing speed at all levels of the orthographic system. The processing of linguistic information is tuned to the frequencies of many types of association.

In vocabulary learning, frequency effects are readily apparent. High frequency words tend to be learned before low frequency words (e.g., N. Schmitt, D. Schmitt, & Clapham, 2001) and words encountered more often in a text tend to be learned before less frequent words (e.g., Waring & Takaki, 2003; Webb, 2007). Therefore, knowledge of language learners' vocabulary sizes, estimated by tests such as the Vocabulary Size Test (Meara & Jones, 1990), can provide insight as to their overall language ability and can be useful in monitoring their vocabulary learning over time.

Similarly, frequency affects the extent of learning formulaic language encountered during reading while listening (Webb, Newton, & Chang, 2013) as well as from reading a single text than through multiple texts (Webb & Chang, 2015). It can be seen that language development is, therefore, very much driven by frequency effects.

Continuing to focus on frequency effects, Malone (2018) explored how exposure frequency, aural enhancement of input, and working memory affected incidental vocabulary learning. In his study, 80 English learners were given reading tasks with target words embedded two or four times. The measures were used for triangulation of data following vocabulary posttests. The researcher found that measurable learning occurred at both frequency exposure levels, a finding consistent with previous studies.

Transfer of Learning

Transfer of learning is a concept that is central to education and is the ultimate aim of teaching and learning.

> There is no more important topic in the whole psychology of learning than transfer of learning . . . . Practically all educational and training programs are built upon the fundamental premise that human beings have the ability to transfer what they have learned from one situation to another . . . . The basic psychological problem in the transfer of learning pervades the whole psychology of human training . . . . There is no point to education apart from transfer. (Desse, 1958, p. 213)

Transfer has been discussed in the fields of psychology, education, and management and has been defined in several ways. It has been described as occurring when people carry over something they learned in one context to a significantly different context (Fogarty, Perkins, & Barell, 1992) or the application of knowledge learned in one setting or for one purpose to another setting and/or purpose (Gagne, Yekovich, & Yekovich, 1993). In management, transfer is defined as the effective and continuing application of knowledge and skills gained in training to the job context and maintained over a period of time on the job (Broad & Newstrom, 1992). Educational researchers have further stated that transfer of learning is a core belief of educators and that knowledge will be learned, retained, and applied in appropriate situations (Ripple & Drinkwater, 1982, p. 1947), that evidence of learning is a person being able to demonstrate that learning later (Perkins & Salomon, 1996, p. 423), and that it is the process and the extent to which past experiences (transfer source) affect learning and performance in a new situation (transfer target) (H. Ellis, 1965).

Transfer of knowledge goes beyond regurgitating memorized material. From the definitions given above, it is the application of previously acquired knowledge and experiences to new concepts or situations. Transfer has been described as near or far depending on how similar the contexts are. However, the notions of near and far transfer are difficult, if not impossible, to measure (Perkins & Salomon, 1992). Perhaps intuitively, teachers will try to teach to a test (i.e., striving for near transfer), meaning that they will teach their students in a way that matches the content and format of a given test (Swain, 1985). This approach has been particularly observed for high-stakes tests, such as when performance determines acceptance into a university. Buck (1988) noted that both teachers and students tend to use classroom activities that directly address test demands (p. 17). This influence of tests on teaching, the washback effect, is a clear parallel of making the learning context match the transfer context.

Thorndike and Woodworth (1901) introduced the notion of transfer of practice. They described how improvement in one mental function could influence a related one. For example, how training in the estimation of the area of various rectangles can improve the estimation of area of different shapes. This notion was also raised by Morris et al. (1977) through their criticism of the Levels of Processing framework (Craik & Lockhart, 1972) where the degree of similarity between the guided and recall tasks determines the success of the recall task, also referred to as the learning and transfer contexts (Thorndike & Woodworth, 1901). This criticism was the basis of Transfer Appropriate Processing (Morris et al., 1977). However, there is a notable caveat: it is not the similarity of the tasks that is a key determinant of transfer; rather, it is the perceived similarity of the tasks from the viewpoint of the learner (Gick & Holyoak, 1987).

It is not in the scope of this study to delve into the numerous definitions, types (Positive and Negative, Simple vs. Complex, Near and Far, Automatic and Mindful), or challenges of transfer. The terms are rarely defined consistently and there is little agreement how to classify different tasks and domains. That being said, in this study, transfer of knowledge is defined as vocabulary studied in one course being used in a writing task in another course.

The Role of Output in Language Learning

Having reviewed several theoretical discussions concerning learning and memory, it is important to further examine the role of output in language learning. As mentioned in the previous section, Tomasello (2000) argued that language development is contingent upon language use, part of which lies in output.

Two terms central to the role of output in language learning are *pushed output* and *generation*. Briefly, pushed output serves as an attention-focusing device where, as a result of being pushed to produce language, learners notice the gap between what they intend to convey and how it is actually received (Leeser, 2008). When combined with feedback subsequently received, forming an output-input cycle, learners can be guided to convey meaning more accurately (see below for further discussion of pushed output). Generation, or more specifically the theory of Generative Learning, proposes that learning occurs as a result of associations being formed between new information and pre-existing knowledge, or as learners construct meaningful understandings of new information (Meyers, 2013) (see below for further discussion of Generative Learning).

Pushed Output

In Swain's (1985, 1995, 2005) output hypothesis, the production of language (i.e., output) is considered a significant contributing factor to language acquisition. More specifically, Swain posited that the type of output that is necessary to promote second language development is pushed output where learners have to develop certain communication strategies in order to be understood, trying out new grammatical structures or vocabulary that have yet to be mastered. However, it must be noted that the pushing seems to cease once the speaker is understood by the listener. Swain (1985) found that once there was no longer a perceived hindrance to communication, efforts at being more comprehensible ended. Building on her original hypothesis, Swain proposed that output possesses the following functions: noticing/triggering, hypothesis testing, and metatalk.

Support for pushed output also came from Browne (2002), who compared pushed output with extensive reading and activity-based tasks with regard to second language vocabulary acquisition. The purpose of that study was to evaluate the effectiveness of two common second language vocabulary teaching techniques and compare them with a technique based on the pushed output hypothesis. The reading task consisted of a passage which contained 10 target words. The activity-based tasks were opportunities to practice the target words in crossword puzzles, matching activities, and word searches. The pushed output task required the students to use each of the target words in at least two complete sentences. Posttests indicated that the pushed output task resulted in significantly more new words being learned than in either the reading or activity-based tasks. Browne argued that the reason that the pushed output task was more effective was because it required the deepest level of processing, thus supporting Craik and Lockhart's (1972) Depth of Processing Theory.

While Browne (2002) compared three different tasks, De la Fuente (2002) focused on the effects of negotiation in the oral acquisition of L2 Spanish vocabulary. The participants in the study were randomly placed in three groups: nonnegotiated with premodified input (NNPI), negotiation of input without output (NIWO), and negotiation of input plus output (NIPO). The target words were 10 Spanish words that the participants had not been exposed to previously. The target words were people or objects that were placed in a room. All groups were given the task of finding the target objects. The NNPI group listened to instructions and could not ask questions. The

NIWO group listened to the same instructions but were allowed to ask as many questions as needed in Spanish. Both the NNPI and NIWO groups performed the task a second time with the order of items changed to avoid a task effect. The NIPO group followed the same instructions as the NIWO group with one change. In the second performance, the participants became the providers of information and had to give instructions to native Spanish speakers, therefore, utilizing a pushed output component. The results indicated that, while negotiated interaction positively affected overall L2 comprehension, only negotiated interaction with pushed output promoted both receptive and productive vocabulary acquisition. As the participants in the NIPO group had to produce the target words to give the same instructions, this resulted in a deeper level of processing that reinforced learning of the target words. Clearly, the studies that showed the effect of pushed output on vocabulary learning (e.g., Browne, 2002; De la Fuente, 2002) have clear links to the Levels of Processing framework (Craik & Lockhart, 1972).

For the purposes of this study, the noticing function is the most pertinent, as it proposes that learners might notice a gap in their linguistic ability while expressing meaning. Swain (1998) suggested that noticing these gaps could result in learners making the conscious decision to address these linguistic deficiencies by focusing on conveying a message accurately, coherently, and appropriately. Through this process, it can be seen how individual tasks push learners to compare their output with both the model text and that of their peers. In turn, this provides opportunities for the learners to seek out appropriate linguistic forms that allow them to more accurately express meaning. Indeed, Leeser (2008) found that learners who were pushed in their output via text reconstruction tasks showed a decrease in non-target-like forms and an increase in producing output attempts at matching the forms in the input, which lends support to Swain's (1995, 1998, 2005) output hypothesis.

When language learners speak or write in an L2, they are likely to pay attention to language form even without teacher intervention (Byrnes & Manchón, 2014). Writing, in particular, affords learners the opportunity to examine and monitor their language because it does not require the same immediacy as speech (Kim, Lowenstein, Pearson, & McLellan, 2001). As a result, this opportunity to notice gaps between their intended message and ability to deliver the message accurately, coherently, and appropriately can lead to learners correcting their own writing.

Generation, Generative Learning, and Generative Use

Generation refers to the connections a learner builds between the different elements of new information within itself, and between the new information and the learner's pre-existing knowledge (Wittrock, 1974). In effect, it is the act of making sense of new information through reorganization and integration in order for this information to be used in new situations. This is the premise of the theory of generative learning. Wittrock's (1974) generative model of learning stems from previous work (e.g., Bartlett, 1932; Katona, 1940; Piaget, 1926; Wertheimer, 1959) that conceptualizes learning and memory as being constructive acts. Furthermore, this work posits that there is a distinct difference between learning for memorization and learning for understanding. Several researchers have described generative learning theory in similar ways: that generative learning is the tendencies of people to generate perceptions and meanings that are consistent with their prior learning (Osborne & Wittrock, 1985); that learning is the result of generating associations between new and pre-existing information, and between individual new items of information (Meyers, 2013); and that generative learning is the process of transforming incoming information into usable knowledge (Fiorella & Mayer, 2016). These researchers showed how generative learning describes the method by which people construct knowledge through the linking of stimuli and stored information.

Generative use is specifically described as when learners encounter or use words in a context that is different which they previously met them (Nation & Webb, 2011). This description most closely matches Meyers' (2013) definition of generative learning. With an explicit focus on vocabulary learning, generative use has been shown to encourage depth of processing and to increase the likelihood of remembering a word (De la Fuente, 2002; R. Ellis & He, 1999; Joe, 1998).

Other researchers have examined how learner-generated marking can be effective for a variety of assessments. Learner-generated marking is considered an example of generative use as the process involves identifying vocabulary in a different context than when first encountered. Leutner, Leopold, & Den Elzen-Rump (2007) approached this technique from a self-regulated learning perspective. They drew parallels from self-regulated learning to generative learning as both help learners process information and make their own decisions regarding the optimal application of this information to the learning task and learning situation. The participants in the study were placed in three experimental conditions: learning strategy training with self-regulation training, learning strategy training without self-regulation training, and a control without learning strategy training and without self-regulation training. The learning strategy consisted of text-highlighting and the self-regulation strategy had the participants practiced ways of regulating the text-highlighting strategy in a goal-oriented manner, reflecting on their approach. In a multiple-choice immediate posttest, Leutner, Leopold, & Den Elzen-Rump found the participants who had undergone both learning strategy and self-regulation training drew the most benefit with a large effect size (d = 1.16).

Rickards and August (1975) also investigated learner-generated marking but compared it to experimenter-provided marking. The participants were given a passage with one sentence in each paragraph underlined by the researchers, or an unmarked version of the same passage in which they had to self-select the sentences to underline. After the task was completed, the participants had to recall in writing as much as they could about the passage. Results of the posttest showed that the participants who were free to underline sentences in the text were superior to the participants in other groups on the measure of passage recall. In short, researchers have shown that a technique such as learner-generated marking is effective for assessments such as multiple-choice tests (Leutner, Leopold, & Den Elzen-Rump, 2007) and free-recall tests (Rickards & August, 1975) as well as short-answer tests (Amer, 1994; Blanchard & Mikkelson, 1987), fill-in-the-blank tests (Yue, Storm, Kornell, & Bjork, 2015), and essay questions (Annis & Davis, 1978; Rickards & Denner, 1979).

In the context of this study, generative use is the creation of links between stimuli, vocabulary in this study, and storage of this vocabulary. These links must be generated through learning. Therefore, generative use is defined as application of techniques to learn vocabulary. While Osborne and Wittrock (1985) focused on the field of science education, parallels can be drawn to second language acquisition in that researchers must consider how received information is prioritized. In the same way that links are generated between external stimuli and stored information, language learners generate links between new and known vocabulary, thereby incorporating the new information into their existing lexicon. It can be said that generative learning theory is concerned with connecting words to one another. Therefore, this theory is particularly relevant in examining its effects on productive use of vocabulary.

Hilton (2008) investigated the contribution of lexical knowledge to spoken fluency through the use of a monologue-type task in which participants described a video immediately after viewing it. She found that there was a clear relationship between known L2 vocabulary and productive use. The implication being that vocabulary and grammatical knowledge facilitate more spoken language production for this type of task. One limitation of Hilton's study was that it did not assess the participants' interactional competence due to the monologue nature of the task. However, the findings of her study can be linked to L2 vocabulary use in writing.

Similarly, Coxhead (2008) also found that the type of vocabulary knowledge contributed to the likelihood of target word use. The types of vocabulary knowledge include knowing how to use a word in writing, the context of the writing, and the topic of the writing. In addition, she discovered that perceived risk or confidence in using words in writing and knowledge of the topic were significant determinants in usage of the target vocabulary. Some of Coxhead's participants reported that they tried to avoid making errors when using the target words and others reported that they did not use particular words due to a lack of confidence in their ability to do so correctly. Similarly, familiarity with a topic led to the participants feeling more competent or able to use those words in writing about that topic.

Linking Pushed Output and Generation

Like pushed output, generation has been identified as one of the deepest levels of processing and linked to the Levels of Processing framework (Stahl & Fairbanks, 1986). Several researchers have investigated different methods of retelling, including the 4/3/2 approach, which decreases amount of time for the retelling (Arevart & Nation, 1991; Maurice, 1983), and the Read and retell approach (Simcock, 1993). While there are methodological differences between the two forms of retelling, the teacher or researcher looked for the use of specific vocabulary in the retelling activity. The key in linking these studies to LOP lies in the central role generative use plays in the retelling task.

In an effort to investigate the extent to which pushed output had an effect on vocabulary learning without the specific intention to learn vocabulary, Joe (1998) conducted a study where participants were assigned to one of three treatments: (a) reading and retelling a text with explicit generative training with no access to the text during the retelling phase (experimental group), (b) reading and retelling a text without explicit generative training with access to the text during the retelling phase (comparison group 1), and (c) neither reading nor retelling a text (comparison group 2). The participants' explicit generative training consisted of instruction in three things: (1) recalling prior experience and knowledge to make sense of unfamiliar words in the text, (2) to paraphrase, use synonyms, examples, or analogy, and (3) to discuss why some examples of generative strategies were better than others. The training was intended to promote generative learning by leading learners to engage in generative or deeper processing. Upon completion of the treatments, the participants took two multiple-choice tests to measure incidental vocabulary learning. Joe found that words that were used generatively in the retelling were more likely to be learned than when the words were used in the same way that they had appeared on the input sheet (i.e., comparison group 1 outperformed the experimental group). She stated that this result can be explained by the increased learning burden experienced by the experimental group, who had to read the text, decide which words or concepts needed to be learned, as well as to comprehend and generate new and relevant information. In contrast, comparison group 1 was given access to the text, which might have afforded those participants more time to focus more on generating ways of using the target vocabulary meaningfully without the burden of memorizing those words. This result was in line with Swain's (1998) findings involving dictogloss. Swain argued that dictogloss tasks pushed learners' output because they had to produce language that was not their own and was potentially beyond their current proficiency level. Both Joe and Swain asked their participants to engage in a retell task. From a vocabulary learning perspective, the learners had to productively retrieve target vocabulary and use the words generatively.

Second Language Vocabulary Learning

Vocabulary is arguably the most important component in language development. Words enable people to communicate and certain words are more important to learn than others. In fact, in the ongoing effort to determine exactly which words should be prioritized by language learners, Browne, Culligan, and Phillips (2013b) produced a list of 2,801 core high frequency vocabulary that provides more than 92% coverage of general English texts. However, learning vocabulary does not automatically lead to the ability to use it. In this section, I review the currently available measures of productive vocabulary knowledge, the various aspects of word knowledge, and commonly used strategies for second language vocabulary learning.

Receptive and Productive Vocabulary

Laufer and Paribakht (1998) investigated the relationships between what they called passive vocabulary (receptive vocabulary), controlled active vocabulary (words learners can use if required), and free active vocabulary (words learners voluntarily choose to use). The participants were 79 young adults from an Israeli high school studying English as a foreign language (EFL). The point of interest is the distinction they made between controlled active and free active vocabulary, that is, the ability to use the word when required (such as in a task) versus the ability to use the word voluntarily. The researchers found that these three dimensions of vocabulary knowledge developed at different rates, with free active vocabulary developing the slowest and being least predictable because, along with understanding the meaning of words, productive use requires knowledge of grammar and context. Knowledge of individual words falls somewhere along this continuum of receptive to productive vocabulary. However, the exact nature of how learners progress along this continuum of vocabulary knowledge is still unclear.

Citing an increasing interest in vocabulary acquisition and pedagogy coupled with the inconsistent use of vague definitions, Henriksen (1999) sought to provide a more complete and standardized definition of vocabulary knowledge by describing three separate but related vocabulary dimensions: partial-precise knowledge, depth of knowledge, and a receptive-productive dimension. The first dimension of partial-precise knowledge is related to vocabulary size, or the number of words with a primary meaning that a learner has. The depth of knowledge dimension refers to the degree to which a learner knows a certain word in addition to the primary meaning and describes the type of lexical knowledge referred to by other researchers as *lexical competence* (e.g., Nation, 1990, 2001; Richards, 1976). The third dimension concerns the degree of mastery of lexical items and how well learners are able to use them in comprehension and production. Henriksen found that the relationship between the two knowledge dimensions and the acquisition of word meaning has been overlooked due to either (a) the focus of much of the research on measuring learners' vocabulary size or (b) the comprehension and production processes involved in accessing vocabulary at a given point in time. As a result, she concluded that further research is needed to determine how learners' interlanguage semantic networks develop over time.

Several studies have been conducted concerning vocabulary knowledge and productive use. Hilton (2008) examined the contribution of receptive lexical knowledge to spoken fluency using an oral corpus of second language (L2) English, French, and Italian that was constructed to identify the linguistic characteristics of different L2 proficiency levels. The 56 non-native participants —33 studying English as an L2, 11 studying Italian, and 12 studying French— were required to describe a short video sequence immediately after viewing it. The nature of the task meant that interactional competence could not be assessed and the focus was instead on aspects of lexical encoding. Fluency was determined by computer analysis of the participants' speech rate, mean length of run, hesitation, percentage of production time spent hesitating, average hesitation time, and rates of hesitation and retracing. The results showed that greater vocabulary knowledge correlated positively with fluency as measured by speech rate (r = .58). The results suggest that greater vocabulary knowledge can lead to more fluent language production. The study shows the importance of lexical development in L2 classrooms.

More recently, Uchihara and Saito (2019) investigated if productive vocabulary knowledge could predict oral ability. Using Lex30 (Meara & Fitzpatrick, 2000) to estimate their participants' productive vocabulary, the researchers utilized a series of picture description tasks to determine comprehensibility, accentedness, and fluency. While productive vocabulary scores significantly correlated with L2 fluency at r = 0.34, they did not find significant associations with comprehensibility or accentedness. These results suggest that learners with a larger productive vocabulary tend to speak more fluently and, along with Hilton's (2008) findings, has ramifications for writing tasks such as those done for this study.

Continuing this line of inquiry, Stæhr (2008) investigated the relationship between vocabulary size and the language skills of reading, writing, and listening. The participants in his study were 88 Danish EFL learners in lower secondary education. Stæhr employed the Vocabulary Levels Test (VLT) developed by N. Schmitt et al. (2001) to assess knowledge of the most frequent 2,000 word families. Stæhr's results supported previous research, as he found strong, statistically significant relationships between receptive vocabulary knowledge and reading, writing, and listening ability. He stated that 52% of the variance in participants' writing scores were accounted for by vocabulary size.

Miralpeix and Muñoz (2018) further examined the relationship between vocabulary size and the EFL skills of reading, writing, listening, and speaking. With the inclusion of speaking, their findings supported and expanded on Stæhr's. Along with the finding that receptive vocabulary size correlates with overall language proficiency, Miralpeix and Muñoz found that writing had the strongest correlation out of the four skills with vocabulary. They conducted a regression analysis and the results indicated that receptive vocabulary size predicted approximately 30% of the variance in writing ability. Both Stæhr's and Miralpeix and Muñoz's studies underscore the importance of vocabulary size for language proficiency. Specifically, knowledge of the most frequent 2,000 words in English was critical for language learners.

While several studies have shown the relationship between vocabulary knowledge and writing, others have examined how vocabulary use in writing can be influenced. Kaur and Hegelheimer (2005), building on research by Stevens (1995) that showed the benefits of contextualized versus decontextualized vocabulary learning, suggested that the use of an online concordance program results in transfer of vocabulary knowledge to an academic writing task and highlighted how concordancers provide additional contextual information that aids learning. However, this study had several limitations: the writing was done outside of class so time on task and use of dictionaries or the concordance program could not be controlled, instruction of the target vocabulary was lacking, and the list of target vocabulary was available during the writing task. Similarly, other studies have investigated the influence of concordancing tools in writing (e.g., Chan & Liou, 2005; Horst, Cobb, & Nicolae, 2005; Kennedy & Miceli, 2010; Sun, 2007). However, these studies focused on the use of specific vocabulary-learning tools and not directly on the transfer of learned vocabulary to writing regardless of learning strategies employed.

Other researchers have indirectly investigated vocabulary transfer in writing tasks (e.g., Laufer, 1998; Laufer & Nation, 1999; Lee, 2003). Laufer (1998) measured the gains of passive, controlled active, and free active vocabulary for one academic year and found that the free active vocabulary use in writing did not progress despite gains in the scores of passive and controlled active vocabulary. Laufer and Nation (1999) examined learners' writing based on the assumption that the absence of target vocabulary implies lack of confidence which in turn reflects imperfect knowledge. Lee

(2003) found that the effects of explicit vocabulary instruction led to approximately 64% of target vocabulary being used in writing compared to only 13% after reading comprehension exercises. In the studies cited above, Laufer (1998), and Laufer and Nation (1999) assessed learners' free active vocabulary as measured by the Lexical Frequency Profile (LFP). However, the LFP analyzes the amount of words that fall in certain frequency bands and not specific target vocabulary. Any conclusions drawn regarding transfer of vocabulary from these studies are, therefore, implicit. While Lee (2003) examined the occurrence of target vocabulary in addition to LFP analyses, her study had a very narrow scope in that the 36 target lexical items were obtained from an adapted article on bullfighting, the vocabulary instruction occurred in a single 75-minute lesson, and the learners were instructed to write on a specific topic which arguably imposed certain constraints on the learners. Furthermore, all the measures, instruction, and tasks were administered in the same context—the vocabulary learning and writing were conducted in the same class with the same teacher—which might have alerted the participants to the true purpose of the study.

Aspects of Word Knowledge

Vocabulary knowledge has varying degrees of granularity. For instance, one can have knowledge of the written form of a word, its grammatical position, collations, appropriateness, and meaning, each of which has their own learning burden. The principle of learning burden is that the more a word a word reflects patterns and knowledge that learners are already familiar with, the easier it is to learn the word (Nation, 2001). Nation further indicates that the closer or more related an L2 is to a learner's L1, the easier it will be to learn L2 words. It is also important to consider whether the vocabulary knowledge and use are receptive or productive. Table 2 lists the aspects of word knowledge from Nation (2001). The aspects of word knowledge most related to this study are Form-written, the three components of Meaning, and the three components of Use, all in receptive (R) and productive (P) knowledge. Therefore, if learners know a word receptively, this is a general statement about the aspects of knowledge a learner has and how it pertains to listening and reading (Nation, 2001). For learners to transition from receptive to productive knowledge, different learning activities need to be utilized for the most effective learning to take place. Table 3 (Nation, 2001) outlines different kinds of lexical knowledge (detailed in Table 2) and effective kinds of learning and activities.

Reports from language learners corroborate this framework. Coxhead (2008) investigated how L2 writers select vocabulary for productive use from a source text. They were not given instruction on the words prior to this task. Her participants reported that they were primarily concerned with the written form of the words and word meaning. Other major concerns included referents and associations, which reflected the participants' belief that the degree of vocabulary knowledge significantly influenced word use in writing. In summary, Coxhead's study supported Nation's (2001) conceptualization of the aspects of word knowledge by illustrating how L2 learner are concerned with the different kinds of lexical knowledge.

Second Language Vocabulary Learning

In this section, I provide an overview of word lists and word cards as common second language vocabulary learning strategies. I review some of the literature examining the effectiveness of word cards on L2 vocabulary learning followed by reasons why relying solely on word cards might not be beneficial for vocabulary learning or overall L2 development.

Table 2. *What is Involved in Knowing a Word*

Knowing a word	Components		Aspects of knowledge
Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency ...)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Note. R = receptive knowledge, P = productive knowledge. Adapted from *Learning Vocabulary in Another Language* (p. 27), by I. S. P. Nation, 2001, Cambridge, England: Cambridge University Press. Copyright 2001 by Cambridge University Press. Used with permission.

Table 3. *Kinds of Vocabulary Knowledge and the Most Effective Kinds of Learning*

Kinds of knowledge		Kinds of learning	Activities
Form		Implicit learning Involving noticing	Repeated meetings as in repeated reading
Meaning		Strong explicit learning	Depth of processing through the use of images, elaboration, deliberate inferencing
Use	Grammar Collocation	Implicit learning	Repetition
	Constraints on use	Explicit learning	Explicit guidance and feedback

Note. Adapted from *Learning Vocabulary in Another Language* (p. 35), by I. S. P. Nation, 2001, Cambridge, England: Cambridge University Press. Copyright 2001 by Cambridge University Press. Used with permission.

Word lists and word cards. Learning vocabulary using word lists and word cards are strategies that are widely used by language learners (Nakata, 2011; N. Schmitt, 1997; Wissman et al., 2012). Word lists are sheets of paper with a list of L2 words and accompanying L1 translations, and word cards are a set of cards with an L2 word on one side and, typically, the L1 translation or an L2 definition (or both) on the other side. Initially, the two seem similar in that they present the same information—L2 words and L1 translations—but there are key differences that result in different degrees of retention. The first lies in the ability to change the order of the items; the order of the items in a word list cannot be changed, while it can be changed with word cards. Another difference is the presence of retrieval practice. For word cards, learners have to actively recall the L2 word form or meaning as they are presented on different sides of the card. However, word lists display both at the same time removing the need for retrieval practice, which has been shown to produce superior retention (Mondria & Mondria-de Vries, 1994; Nation, 2001; N. Schmitt & D. Schmitt, 1995; Waring, 2004).

The terms *word cards* and *word lists* have been conflated with other terms. These include *list learning* (Griffin & Harley, 1996), *paired associate learning* (Carroll 1963; Higa, 1965), and *learning word pairs* (Nation, 1982). All these terms that have been used interchangeably to mean matching L2 words with L1 translations or definitions.

Empirical studies have shown that word cards and word lists are both effective and efficient methods of learning vocabulary (Elgort, 2011; Fitzpatrick, Al-Qarni, & Meara, 2008; Laufer, 2003b; Nation, 2001; Webb, 2009). Word cards, in particular, result in superior retention when the following guidelines are followed. The nature of word cards facilitates retrieval as the word form and meaning are presented separately (Baddeley, 1997; Royer, 1973). Pairing L2 words with their L1 translations has been shown to be more effective than other options for word card construction such as L2 definitions, L2 synonyms, pictures, or L2 contextual definitions (Laufer & Shmueli, 1997). Elgort (2007) showed that spaced rehearsals using word cards led to her participants acquiring formal, semantic, and procedural knowledge of the words studied similar to that of native speakers. The participants also achieved native-like speed of lexical access in a range of priming conditions. Elgort concluded that learning from word cards contributes to implicit knowledge, that is, normal language use. Studies have also shown that vocabulary learned in this manner is both resistant to decay (Fitzpatrick et al., 2008; Thorndike, 1908) and retained over several years (H. Bahrick, L. Bahrick, A. Bahrick, & P. Bahrick, 1993; H. Bahrick & Phelps, 1987).

Why word cards might not promote acquisition. Despite word cards being commonly used for learning a large number of words in a short time, there are concerns that the reasons for using word cards are not aligned with those of language development. Cobb (1999) argues that using word cards or lists creates superficial knowledge which makes it difficult to transition to spontaneous language production. In a study examining how computerized learning activities can utilize the benefits of list learning and natural reading while eliminating their shortcomings, Cobb constructed a corpus of his participants' reading materials and identified 2,387 of the most common words. The participants used this corpus to study the target vocabulary which produced a concordance of all the word's occurrences in the reading materials, then by showing a selected target word in the source text. The results showed that vocabulary learning using word lists produced strong short-term gains. However, long-term retention was weak and the participants were limited in their ability to apply the words in gaps in a novel text. In contrast, participants who used the corpus were more successful in the transfer of comprehension to novel texts in both the short and long term. Cobb posits that the vocabulary learning using word lists was weaker due to the fragile lexicons that were produced as a result of the one-to-one connections of L2 word and L1 translation/definition. This conclusion is supported by other researchers who report that the word list technique does not improve learners' reading comprehension or help with future word learning (Miller & Gildea, 1987; Nesi & Meara, 1994). However, if the purpose of vocabulary study is merely to memorize large quantities of words, then the minimal nature of paired associate learning does this quickly (Fitzpatrick et al., 2008; Nation, 1980; Thorndike, 1908).

According to Nakata (2008), list learning is a kind of rehearsal, as it is "an activity to encode new information into long-term memory through overt or silent articulation" (p. 5). Regardless of whether learners employ word cards or word lists, spacing effects apply (see Crammed vs. Massed vs. Spaced Learning above). Elgort's (2007) study supports this view. However, if learners tend to cram for tests which means that they would not benefit from spacing effects which are contingent upon frequent rehearsals with the appropriate distribution of rehearsal opportunities.

"Instantiation involves recalling or experiencing a particular instance or example of the meaning of a word," though it is possible to learn vocabulary without instantiation as learners might make L2 word-L1 translation connections without thinking of a particular instance of the meaning (Nation, 2001, p. 70). In other words, *instantiation* might involve *generation*, but generation is not necessarily present during instantiation (Nation, 2001). It has also been suggested (Anderson, Stevens, Shifrin, & Osborn, 1978) that instantiation is linked to generative use and that using vocabulary communicatively contributes to learning because each word is linked to a meaningful example and not merely its L1 translation.

Another conundrum regarding learning through word cards is efficient use of time. Specifically, is it better for learners to spend time making their own word cards or using pre-made ones? Wilkinson (2017) interviewed seven Japanese university students regarding the perceived usefulness of learner-prepared word cards as a method of deliberate vocabulary learning. In his analysis of the interview transcripts, the participants revealed that making their own word cards were an effective method for vocabulary study. In addition, despite the additional time required to prepare their own word cards, the participants considered it a quick way to learn new words. However, it must be noted that the participants were not asked to directly compare self-made and pre-made word cards. As such, it is possible that different learners will have contrasting responses.

Wong (2017) conducted case studies of six EFL vocabulary learners regarding their vocabulary learning beliefs and strategies. Several participants expressed a preference for self-made word cards, yet they also reported a lack of time and/or willingness to create them. In general, word cards are a quick and effective way to study vocabulary. Nation and Webb (2011) suggested that learners can remember between 30 to 70% of the meanings on the first run after making word cards. However, making word cards is a time-consuming affair. If a learner was to make 100 word cards containing only L1 translations, that might take more than an hour.

Computer-based word cards form a natural evolution from pen-and-paper versions. They relieve much of the burden of creating word cards, changing the sequencing, and adjusting the spacing of the learning is removed from learners. Flashcard websites and their smartphone versions are able to perform these functions, which in turn might help students learn more effectively compared to paper-based counterparts. These computer-based tools adapt to a learner’s performance and ensure that more time is spent learning unknown or difficult items. Several researchers have argued that these digital tools offer benefits over traditional paper-based word cards (N. Ellis, 1995; Hulstijn, 2001; Nation, 2001). Another benefit to computer-based word cards is that many of the currently available resources are designed to reduce the amount of practice needed for items in which the user has demonstrated mastery. Miyatsu, Nguyen, and McDaniel (2018), in their analysis of five popular study strategies, explained that students lacked the requisite metacognitive accuracy in determining when to remove cards from a stack for future study which resulted in two problems. First, cards were removed before they were successfully learned. Second, the removal of these cards prevented the learner from encountering them again unless they were deliberately reinserted.

The effectiveness of computer-based word card use compared to paper-based word cards and word lists has also been investigated. Nakata (2008) conducted the first study to empirically compare computer-based word cards to paper-based ones and word lists based on several benefits lacking in cards and lists. Specifically, computer algorithms are able to implement expanded rehearsals and retrieval practice while varying the presentation order of the items thereby avoiding the problems of list effect. He sought to determine which materials led to the optimum spaced learning. His participants were 217 Japanese high school students, with approximately one-third of the total group in each condition. All were told to study 10 English words. The findings from this study were mixed. Nakata found that participants using computer-based word cards outperformed those using word lists in the delayed posttest scores, $F(2, 180) = 3.23$, $p < .05$. However, there was no statistically significant difference between the computer-based cards and the paper-based cards $F(2, 194) = 0.39$, n.s.. As a possible explanation for this finding, Nakata suggested that the participants were not trained in using strategies for spaced learning or were unfamiliar with Computer Assisted Language Learning. This point is relevant to this study as the two self-study methods of Quizlet (2017) and Moodle (2016) are computer-based.

Since Nakata's study was conducted, several Internet-based services such as Memrise (n.d.) and Quizlet (2017) have been introduced that now make it easier to study vocabulary using computer-based word cards. These services allow learners to create word card sets, store them on the Internet, and use them on their smartphones. These applications adapt to the learners and change item sequencing or spacing so as to make the study as efficient as possible while removing the pitfalls of having learners decide when and if cards should be removed.

As digital learning resources have evolved and become more widely used, they have prompted ongoing investigations to their effects on learning. Many now have gamified features, potentially making their use less tedious and fostering a more positive attitude on the part of users. Sandberg, Maris, and Hoogendoorn (2014) examined the effect of these features and noted that gaming contexts and intelligent adaptation contributed to improved learning. Furthermore, these resources are also intended for use on smart phones, and this feature allows learners more freedom in deciding how and when to study. Self-pacing has been found to have significant, positive effects on vocabulary learning (Huang, Yang, Chiang & Su, 2016; Wang & Shih, 2015). In short, as technology improves and becomes more ubiquitous, newer ways to learn are constantly being developed that improve learning effectiveness while also increasing learning enjoyment. Yet, it must be remembered that word cards as a tool for acquisition still possesses some limitations.

Components of Good Writing

In this section, I provide several definitions of *good writing*, examine the contributions of lexical knowledge to writing quality, and show how Zipf's law illustrates the importance of high frequency vocabulary.

One of the central issues in teaching L2 writing is determining a definition of *good writing*. Several studies have assessed writing quality on a variety of measures including the use of cohesive devices such as connectives or word overlap between sentences or paragraphs in a text, word repetition, or ratio of pronouns to nouns (Crossley, Kyle, & McNamara, 2016), syntactic complexity (Crossley & McNamara, 2014), lexical richness (Laufer, 1994), lexical complexity (Ong & Zhang, 2010), or lexical sophistication (Guo, Crossley, & Mcnamara, 2013). Due to the variety of factors affecting writing quality, Casanave (2004) noted that there are many diverse and controversial opinions regarding the best approaches for improving writing. Issues include whether fluency or accuracy should be prioritized, and the types of feedback that lead to improvements in writing quality. In short, defining what constitutes good writing remains difficult.

Further complicating the matter is the area of inquiry called Contrastive Rhetoric (Kaplan, 1966). Contrastive Rhetoric is concerned with the differences in writing styles—organizational style, use of logic, and relative relationship between the writer and the reader—across cultures. For language learners, this assumption implies that differences in linguistic features between their first and second languages can cause difficulties when learning the L2 (Casanave, 2004). This means that their L1 will interfere with their L2 and can explain why language learners make certain choices that are considered inappropriate by native speakers. Therefore, vocabulary use in writing is concerned with not only meaning-appropriate use but also genre-appropriate use (i.e., relevant to a particular discourse community). In academic writing, success is judged on adherence to the writing conventions of the institution as well as the field of study (Schneider & Fujishima, 1995). This poses problems for language learners who are likely unfamiliar with the writing conventions of the target culture.

Several researchers have suggested one significant component of writing quality. Astika (1993) performed a multiple regression analysis on the scores of 210 participants to identify the main predictor of writing scores. The participants each wrote two writing samples, one in Spring 1989 and the other in Fall 1990. Two raters evaluated each writing sample using the ESL Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). The analysis revealed that vocabulary accounted for the largest amount of variance (83.75%), more so than content, organization, language use, and mechanics. This result indicated that vocabulary was the main predictor of writing scores as assessed using the ESL Composition Profile.

Focusing specifically on the perspective of educators, Santos (1988) and Jordan (1997) collected data from university professors and instructors. Santos examined the reactions of 178 professors to two academic compositions, one written by a Chinese student and the other by a Korean student. The professors were asked to rate the compositions on content and language. Analysis of the ratings indicated that, while writing errors did not affect comprehensibility, they were considered academically unacceptable, with lexical errors rated as the most serious. Similarly, Jordan (1997) surveyed the instructors of L2 postgraduate students in the UK and analysis of the responses showed that the three problems associated with written work was style (92%), grammar (77%), and vocabulary (70%). These studies underscore the importance of vocabulary, particularly appropriate vocabulary usage, in determining writing improvement.

While vocabulary is certainly one of the cornerstones of good writing, it is also important to consider that the type of vocabulary use will vary according to the context. Researchers have compiled lists of frequently used vocabulary, such as the General Service List (West, 1953), the Academic Word List (Coxhead, 2000), the New General Service List (Browne, Culligan, & Phillips, 2013b), and the New Academic Word List (2013a). Other researchers have analyzed the vocabulary used for various purposes or domains such as language proficiency tests (e.g., the TOEIC Service List, Browne & Culligan, 2016), television (Webb & Rodgers, 2009), or specific academic disciplines (e.g., agriculture research, Martínez, Beck, & Panza, 2009). Even in the area of English for Specific Purposes (ESP), sub-domains representing different registers have been identified. Examples of these sub-domains include English for academic purposes, English for science and technology, English for the workplace, business English, legal English, aviation English, English for medical purposes, English for nursing, and English for research publication purposes (Paltridge & Starfield, 2013). In short, the topic of a writing task plays a significant role in the type of vocabulary use. Topics that are more familiar to the writer, such as writing about their hometown, would produce a certain type of vocabulary compared to less familiar topics such as writing about the effects of global warming on local flora and fauna.

Another variable that determines the types of vocabulary use is the purpose of the writing task. For example, language learners writing in a personal diary are under less stress and scrutiny than when writing to demonstrate comprehension of complex theories. In sum, vocabulary is central to good writing; however, the writing task plays a role in determining the type of vocabulary that is required. For the purposes of this study, writing quality is defined as consisting of well-organized content with a central theme, appropriate use of a wide range of vocabulary, grammaticality, and length.

How Lexical Knowledge Adds to Writing Quality

Having established that vocabulary is one of the main predictors for writing scores, it is then important to examine how lexical knowledge can be improved. The three aspects of knowing a word—form, meaning, and use—as well as how they are applied receptively and productively are complex (see Table 2 above). When considering how word knowledge contributes to writing competence, relevant areas include: the spelling of the word, the word form (grammar and meaning), associated knowledge (included in the concept, referred to by the concept, other related words), grammatical patterns, collocations, and constraints on use. Participants in Coxhead's (2008) study were interviewed regarding their use of target vocabulary and their reports support this framework. All the participants expressed the need to know a word well in order to use it in writing with four out of seven reporting that spelling was a primary concern. Other participants commented that they focused on other aspects of words such as word family members, meaning, appropriate vocabulary choices, and correct word associations.

Furthermore, the various aspects of word knowledge are learned in several ways. Depending on the areas that warrant more attention, or are identified as in need of improvement, learners at different levels of language proficiency need to focus on specific activities that are most effective at producing the desired result. Coxhead (2008) examined the effects of an attention-raising method (i.e., by underlining the target vocabulary in the text provided to students) on target vocabulary use. She found that tasks that ask participants to use target vocabulary resulted in more target word use than tasks which did not. However, upon further investigation, her findings showed that while underlining words potentially led to their use in writing tasks, the participants required, at a minimum, partial knowledge of those words. Words that were underlined but completely unknown were deemed too risky by the participants to be used due to lacking the confidence to use them despite their apparent importance.

The study by Coxhead (2008) suggests that technical or academic vocabulary can be challenging to use as these tend to be more difficult to master. A study of relevance to this discussion was conducted by Kim and Crossley (2018), who developed a model of L2 writing quality using a structural equation modeling approach to analyze a publicly available corpus of Test of English as a Foreign Language (TOEFL) writing samples, the TOEFL iBT Public Use Dataset (Educational Testing Service (ETS), 2019). The model was based on a corpus constructed from the responses of 480 test-takers on source-based and independent writing tasks. The results indicated that 81.7% of the variance in L2 writing quality was explained by lexical decision reaction time scores ($\beta = 0.932$), lexical overlap between paragraphs ($\beta = 0.434$), and mean length of clauses via lexical decision reaction time scores ($\beta = 0.607$). Overall, these results indicate that higher-rated essays tended to contain more sophisticated words. This finding supports Santos' (1988) observations that genre-appropriate vocabulary is critical to writing quality. Moreover, Kim and Crossley (2018) noted that more sophisticated words require longer response times, greater lexical overlap between paragraphs, and more complex grammatical structures.

Zipf's Law

Zipf's Law states that for any given text, a very large portion of the vocabulary can be accounted for by a few words and, conversely, a very large number of relatively infrequent words account for a small portion of the lexis (Sorrell, 2012). Specifically, Zipf (1949) proposed that the rank of a word r multiplied by its frequency, f, results in a constant value, C, and is expressed by the formula $r * f = C$. This ranking is clearly illustrated when applied to the coverage of the sublists of Coxhead's (2000) Academic Word List (AWL). Coxhead identified a group of 570 words to be common to a wide range of academic text and arranged these words by frequency into ten sublists, with sublists 1 to 9 consisting of 60 words each and sublist 10 made up of the remaining 30 words. Table 4 shows the number of words in each sublist, their coverage of a 3,500,000 running word academic corpus (Coxhead, 1998), and the number of tokens per 350-word page for the texts she examined. Following Zipf's formula ($r * f = C$) and multiplying the rank of the sublist with the coverage percentage, the value C remains constant at 3.6 up to the seventh sublist. The implication for language learners, particularly with regard to vocabulary learning, is that some words are much more frequent than others and, consequently, are more important to learn as they will be encountered and used more often.

Table 4. *Coverage of the Academic Word List*

Sublist rank and number of words		% coverage	Number of tokens per 350-word page
1	60 words	3.6%	12.3
2	60	1.8%	6.0
3	60	1.2%	4.2
4	60	0.9%	3.2
5	60	0.8%	2.7
6	60	0.6%	2.4
7	60	0.5%	1.7
8	60	0.3%	1.3
9	60	0.2%	1.0
10	30	0.1%	0.5
Total	570 words	10.0%	35.1

Note. Adapted from *Learning Vocabulary in Another Language* (p. 190), by I. S. P. Nation, 2001, Cambridge, England: Cambridge University Press. Copyright 2001 by Cambridge University Press. Used with permission.

The Japanese Education Context

Given that the participants in this study were raised in Japan and educated in Japanese schools, a brief overview of the Japanese education context is presented. The Ministry of Education, Culture, Sports, Science and Technology (MEXT, n.d.) has described the Japanese education system as being patterned on the American model with education beginning in kindergartens (*yochien*) for children 3-5 years of age. Students then begin compulsory education in elementary schools (*shogakko*) for six years and junior high schools (*chugakko*) for three years. After completion of the compulsory education program, around 98% of Japanese junior high school students progress to high schools (*koto-gakko*) for another three years (Sugimoto, 2010).

Until recently, English classes started in Japanese public schools in the first year of junior high school. However, in 2011, with the implementation of new guidelines for elementary schools, foreign language communication activities became compulsory for fifth and sixth grade students (Tahira, 2012, p. 4). These weekly activity lessons typically involved singing songs, playing games, or other enjoyable activities that were aimed at nurturing a positive attitude toward communication (MEXT, 2008). In 2020, these activity-based lessons will shift and the teaching of English as a formal subject will begin in upper elementary classrooms, after a two-year transition period (Aoki, 2016). The start of formal English education for the fifth and sixth graders is a result of a government decision to develop more globally-minded citizens and the recognition that the existing system did not achieve the goal of producing students with sufficient competence in debating or negotiating in English (Aoki, 2016).

Problems with Speaking and Listening Development Among Japanese

Although there appears to be a strong emphasis on English language proficiency, Japan ranks comparatively low on various standardized language tests. On the Test of English for International Communication (TOEIC), the national average for Japan was 520 points out of a possible 990 in 2015 (ETS, 2018). For Asian countries, Japan's national average was only higher than that of Macao (490), Mongolia (479), Thailand (478), and Indonesia (464). Japan also has the highest percentage (77%) of test takers who had previously taken the TOEIC, tied with Korea, but outperforms only Afghanistan, Cambodia, Tajikistan, and Laos on the Test of English as a Foreign Language (TOEFL) (ETS, 2016).

There are several possible explanations for this poor performance. First, language study for many students is undertaken only in preparation for university entrance exams. The entrance exam format necessitates rote learning and grammar-translation with minimal, if any, attention on speaking competence (Stewart, 2016). As a result, the students are ill-equipped for the listening and speaking sections on the TOEIC or TOEFL. Second, the students do not receive regular practice involving meaningful, content-filled communication (Disappointing levels of English, 2015). Another contributing factor is the perception by many Japanese that there is no pressing need for English in Japan. Unlike some other Asian countries (Singapore, Malaysia, and the Philippines) where English is a second language, English in Japan is a foreign language, which means that most people in Japan are unlikely to encounter situations that require communication in English (Wakabayashi, 2015). Furthermore, many universities administer the TOEIC to their students several times throughout their university career for placement into language courses, as part of their ongoing preparations for job-hunting, or to use this information in the universities' recruiting of high school students (Murai, 2016). Finally, relative to the local cost of living, there are more Japanese test-takers of exams such as the TOEIC compared to other Asian countries, many of which are still developing their English skills along with their economies. In these Asian countries, proficiency tests can cost a significant portion of the average monthly salary. Therefore, the stakes are higher for those test-takers and they would likely only take the tests when they are confident of their ability.

Pilot Studies

In this section, I describe two pilot studies, Pilot Study 1 and Pilot Study 2. The purposes of Pilot Study 1 were to investigate how the data-collection procedure could be refined in order to ensure the most complete data set possible, examine the results from analysis of data from a smaller sample and use these results to establish the data analysis procedures, and establish the writing prompts for the main study. Pilot Study 2 was conducted to provide insight into the participants' beliefs and reactions regarding the use of the Quizlet (2017) digital word card resource (see Wong, 2017, for details).

Pilot Study 1

Pilot Study 1 Research Questions. Similar to the research questions of the main study, pilot study 1 was designed to answer the following questions:

1. If not instructed to, does vocabulary transfer occur from a vocabulary course to writing courses?

2. Do words that are tested earlier in the semester appear in participants' writing in later dates?
3. Is there a significant difference between students who are reminded and students who are not reminded in terms of this use?

Pilot Study 1 Method. The participants (N = 45, 9 male and 36 female students) in the first pilot study were native first-year Japanese students and were from the same institution as the learners from the main study. The participants were from two intact classes and took a vocabulary course and content course in the fall semester. Vocabulary course classes met once a week and content course classes met twice a week. Instructors in each of the content course classes were briefed on the nature of the study and were provided with copies of an informed consent form to give to the students (see Appendices A and B for the informed consent form used with Pilot Study 1 in English and Japanese, respectively). The students were given time in class to read the consent form and ask any questions they had. I was available to answer any questions the instructors were not confident in answering.

The participants were considered to have low-intermediate to intermediate levels of English proficiency as determined by several measures of English proficiency at tests administered just following entrance to the university in April and at the end of the first term in July. All first-year students take the GTEC (Global Test of English Communication) (Benesse, 2004) in April and the TOEIC (Test of English for International Communication) (ETS, 2017) in July. The GTEC participants' scores ranged from 165 to 289, and their TOEIC scores ranged from 400 to 620. One instructor who taught two class sections agreed to assist with the pilot study. One class was the Reminder (R) group and the other was the No Reminder (NR) group.

However, due to issues in the data collection procedures, which relied on the students resubmitting their essays to the instructor (not to me) after grading, the data sets for 19 participants were incomplete. This left a final *N*-size of 26, with 17 participants in the reminder and nine in the non-reminder conditions.

Target Vocabulary. The target vocabulary was selected by the university. The vocabulary the participants were required to study were a specific subset of words from the New Academic Word List (NAWL) (Browne et al., 2013a). While the entire NAWL consists of 963 words, the complete list was deemed to be too long and too difficult for many of the students to learn to a satisfactory degree during one 15-week semester by program administrators. A smaller list containing 180 words of the NAWL was created by analyzing a corpus generated from the textbooks and teacher-created materials used in the Content Course and then cross-referencing this list with the vocabulary on the NAWL. This process ensured that the 180 words on the smaller list contained only those that the students were certain to encounter throughout the semester. The list of 180 words was further divided into nine sublists of 20 words each. Students were required to study one sublist per week as part of their regular class work.

The three vocabulary tests (see Chapter 3 and the Instruments section below) were required of all students, not just those in pilot study 1. These tests followed the same format as the tests used in the main study (see Appendix C). Briefly, Vocabulary Test 1 included words found on sublists 1 through 3, Vocabulary Test 2 included words found on sublists 4 through 6, and Vocabulary Test 3 included words found on sublists 7 through 9.

Quizlet (2017) study sets and Moodle (2016) quizzes were not prepared for the students in Pilot Study 1. The students were given printed lists of the words and were encouraged to add Japanese translations or transfer the words on the list to words cards, though none, to my knowledge, did this. Observations of the students prior to the vocabulary tests indicated that no students were using Quizlet or word cards before any of the three vocabulary tests were administered.

Instruments and Data Sets. The instrument for the pilot study was the participants' four-genre writing assignments in their content course. The genres assigned during this period consisted of recount, procedure, narrative, and report. The instruction for each of these written genres followed the same series of steps as those of the main study (see Context for the Study in Chapter 2) and the final essays were due in lessons 8, 14, 20, and 28, respectively. The instructor for the course assisted in collecting the writing assignments and I transcribed each essay for analysis.

Procedures. The participants in the Reminder group were told to bring their vocabulary lists to every class meeting. The instructor reminded the students to use in writing at least 5 of the 20 words that they were studying in the Vocabulary Course that week in their Content Course in-class tasks and homework. Table 5 provides details of the Content Course homework with the vocabulary that was studied during the 15-week semester. Other homework and lessons that were not relevant to the genre writing assignments have been omitted.

As can be seen from Table 5, the genre writing assignments were spaced throughout the semester with supporting tasks leading up to each assignment. Schedule variations between the Vocabulary Course and Content Course led to minor irregularities in the introduction of the vocabulary. For the in-class tasks and homework, the students were instructed to use five words from the vocabulary list they were studying that week. The students were not instructed to use five different words for the in-class tasks and the homework. In addition, the use of the five vocabulary items was not monitored by the course instructor.

Pilot Study 1 Data Analysis. In order to determine the target vocabulary use for each group separately, the file was split using the split file command in SPSS using the variable Reminder. This procedure ensured that the Wilcoxon signed-rank test was run for the reminder group and non-reminder group separately.

Analysis of texts. The primary tool for analyzing the collected texts was the Text Lex Compare routine (Cobb, n.d.-a) accessible online. The digitized texts and the vocabulary lists were uploaded and Text Lex Compare showed the degree of lexical repetition across both texts.

Table 5. *Pilot Study 1 Content Course Weekly Homework and Reminder Schedule to Use Vocabulary Previously Studied in Assignments*

Lesson #	Week	Homework	Vocabulary Reminder
1	1	—	—
2	1	(research homework)	—
3	2	(research homework)	—
4	2	Recount Framework	Sublist 1
5	3	Finish Recount Framework	Sublist 1
6	3	Write Recount	Sublist 2
7	4	Finish Recount	Sublist 3
8	4	(research homework)	Sublist 3
9	5	Procedure Framework	Sublists 1-3
10	5	Finish Procedure Framework	Sublists 1-3
11	6	Write Procedure	Sublist 4
12	6	-	Sublist 4
13	7	Finish Procedure	Sublist 4
14	7	(research homework)	Sublist 4
15	8	(research homework)	Sublist 5
16	8	(research homework)	Sublist 5
17	9	(research homework)	Sublist 6
18	9	Narrative Framework	Sublists 4-6
19	10	Finish Narrative Framework	Sublists 4-6
20	10	(research homework)	Sublist 7
21	11	Write Narrative	Sublist 7
22	11	Finish Narrative	Sublist 7
23	12	Report Framework	Sublist 8
24	12	Finish Report Framework	Sublist 8
25	13	Write Report	Sublist 9
26	13	(research homework)	Sublist 9
27	14	Finish Report	Sublists 7-9
28	14	(research homework)	—
29	15	—	—
30	15	—	—

Note. Each sublist refers to one 56-word section of the New Academic Word List.

Writing assignments by genre compared to vocabulary lists. All available writing assignments for each genre were compared against vocabulary lists following the Vocabulary Course study schedule shown in Table 5. This comparison showed target vocabulary usage in each assignment and, more importantly, whether the vocabulary studied for earlier vocabulary tests was used in later writing assignments. Table 6 shows the vocabulary the participants had to study prior to submitting their writing assignments in the content course and the time of their vocabulary test. As the studied words were not tested cumulatively (i.e., words studied for one vocabulary test did not reappear in later vocabulary tests), any persistent use of vocabulary was an indication of learning.

Table 6. *Pilot Study 1 Vocabulary Study and Writing Assignment Schedule*

	Vocabulary Course		Content Course
Week	Sublists	Test	Assignment
1			
2			
3	2		
4			Recount
5		Sublists 1-3	
6			Procedure
7	5		
8			
9		Sublists 4-6	
10			
11			Narrative
12	8		
13			
14		Sublists 7-9	
15			Report

Table 7 shows the comparison outline when examining writing assignments for target vocabulary use. The analysis had two purposes; to identify vocabulary transfer and to potentially reveal evidence of learning. To identify vocabulary transfer, vocabulary currently being studied in the vocabulary course would appear in the writing assignments in the content course. In this analysis, only words originating from sublists 4 and 5 were analyzed. Earlier sublists were not included because these words did not reappear in later tests and participants, therefore, had little incentive to continue studying or remembering that vocabulary. With this in mind, if words studied for a vocabulary test during Week 4 appeared in the writing assignment collected during Week 14, the writer was considered as having successfully learned those words as demonstrated by persistent use.

Table 7. *Pilot Study 1 Comparison Outline for Writing Assignments and Vocabulary Sublists*

Writing assignment	Transferred vocabulary	Learned vocabulary
Recount	Sublists 1-2	Sublists 1-9 (pretest)
Procedure	Sublists 4-5	Sublists 1-3
Narrative	Sublists 7-8	Sublists 1-6
Report	Sublists 9	Sublists 1-9

The column *transferred vocabulary* in Table 7 indicates the words that participants had met in the vocabulary class prior to the content course class genre writing assignment. For the Reminder group, this indicates the sublists from which the participants were supposed to draw vocabulary. The transferred vocabulary is for the sublists that the participants had studied previously in the vocabulary course in preparation for the next vocabulary test, while the learned vocabulary is the sublists used for data analysis that suggests previous knowledge or learning of that vocabulary as determined from the initial writing assignment.

The *learned vocabulary* column shows the lists the assignments were compared to in order to assess whether the participants had learned the vocabulary. A word from sublist 1 that appeared in the Report writing assignment that had previously not been used by the participant might be an indication of learning.

The writing assignments were also compared by genre (e.g., recount compared to report) to examine if there was more target vocabulary use in writing later in the semester. This entailed using Text Lex Compare (Cobb, n.d.-a) to identify the target vocabulary used in all the recount essays and examining the list for any overlap with the report essays. Similar to the analysis used to determine if learning occurred, comparing all the essays in a particular genre (e.g., the recounts) reduced the effect of different topics and writing genres on vocabulary use.

Previous studies (e.g., Santos, 1988) have described how lexical decisions can be considered inappropriate for certain genres or topics and is critical to writing quality. Given that the participants wrote essays comprising four different genres, there would be a degree of variation in type of vocabulary used depending on the genre of the text. By comparing the target vocabulary used in all the essays of a single genre with the essays of another genre, the influence of individual participants on target vocabulary use would be minimized. Overall, the analysis indicated that there was a greater proportion of target vocabulary used from the first essay to subsequent essays.

The final analysis of the texts concerned individual participant's first and fourth essays, the recount and report. It is widely recognized that one participant differs from another in vocabulary knowledge, writing ability, and overall language proficiency. Due to these individual differences, some participants could demonstrate increased vocabulary use while others would not. In an effort to explain changes that are broadly observed from the comparison of genres, participants' recounts were compared to their reports.

The Wilcoxon signed-rank test was used to compare the number of target vocabulary used in the recounts and reports by the R group and NR group. A paired sample *t*-test was not suited because of its distributional assumptions. The Shapiro-Wilk test ($p < .05$) indicated that the assumption of normality was violated. Furthermore, the ratio of the smallest to largest group size was greater than 1.5 (17 participants in the R group, 9 participants in the NR group). Due to the small sample size, the central limit theorem did not apply and normality could not be confirmed. In such cases, Field (2013, p. 228) suggested using the Wilcoxon signed-rank test (Wilcoxon, 1945), which is the non-parametric equivalent of the paired-samples *t*-test. It is used in situations in which two sets of scores are compared from the same participants. The assumptions of the Wilcoxon signed-rank test were checked and met.

Effect size estimates were also calculated in order to show the practical importance of the data. The effect size for non-parametric tests was calculated using the following formula (Field, 2013):

$$r = \frac{z}{\sqrt{N}}$$

where *z* is the *z*-score and *N* is the size of the sample.

Pilot Study 1 Results. Research question 1 asked whether vocabulary transfer occurs from a vocabulary course to writing courses when students are not instructed to do so. This question was the main purpose of the pilot study and was answered by comparing the written texts with the vocabulary lists described above. Table 8 shows the results for the reminder group and non-reminder group.

Table 8. *Pilot Study 1 Transferred & Learned Vocabulary Used by the Reminder (R) and Non-Reminder (NR) Groups for each Genre Writing Assignment*

	Writing	Transferred Vocabulary			Learned Vocabulary		
Group	Assignment	Sublists	Tokens	Families	Sublists	Tokens	Families
R	Recount	1-2	6	6	1-9	14	14
	Procedure	4-5	6	6	1-3	4	4
	Narrative	7-8	20	19	1-6	15	15
	Report	9	35	33	1-9	35	33
NR	Recount	1-2	3	3	1-9	14	14
	Procedure	4-5	1	1	1-3	1	1
	Narrative	7-8	12	11	1-6	10	10
	Report	9	28	26	1-9	28	26

Note. R = Reminder group, NR = Non-reminder group.

Initial comparisons between the Reminder and Non-reminder groups show that the participants who were not instructed to do so used some of the target vocabulary which suggests that vocabulary transfer is occurring between courses. However, the amount of vocabulary use can be influenced through the administration of reminders as demonstrated by the R group. This can be seen by comparing the number of tokens and families in the *transferred vocabulary* column in Table 8. To answer research question 2, d*o words that are tested earlier in the semester appear in participants' writing in later dates?*, and research question 3, *is there a significant difference between students who are reminded and students who are not reminded in terms of this use?*, a Wilcoxon signed-rank test was conducted with the dependent variable being the number of target vocabulary used in the essays.

The Wilcoxon signed-rank test results indicated that the reminder group used more target vocabulary in the fourth writing assignment (E4) (Mean = 3.29) than the first writing assignment (E1) (Mean = 1.65), $T = 90.00$, $p = .017$, $r = 0.58$. The effect size of 0.58 represents a large change in target vocabulary use when reminded. This value is above Cohen's criteria of .50 for a large effect. Therefore, reminding the students to use the target vocabulary led to a significant increase in target vocabulary use from E1 to E4. However, there was a non-significant difference in target vocabulary use in the non-reminder group, $p = .55$. In other words, participants who were not reminded to did not significantly use the target vocabulary.

The results for the reminder and non-reminder groups show that there was a large positive effect when reminded to use the target vocabulary. The focus of this study was not to investigate if something was better than nothing, but rather to observe the degree to which vocabulary study transfers to writing. The students received equal instruction in their vocabulary classes. The reminder that the students in Group 1 received was a reminder to use the vocabulary that they were studying. The results showed that without this reminder, the students in Group 2 did not demonstrate the same degree of vocabulary transfer as observed in their writing assignments.

Pilot Study 1 Limitations. The findings of Pilot Study 1 are limited by several factors. First, the genres and the topics of the writing assignments differed due to the course curriculum. This difference might have influenced the type and frequency of vocabulary use as certain genres and topics might lend themselves to certain vocabulary being used more than others. Second, the in-class tasks and homework were not completed under testing conditions, which means that the students might have referred to the vocabulary lists, dictionaries, or other resources while completing the tasks. Finally, the small *N*-size and incomplete data obtained limit the generalizability of the findings.

While the limitations described above affected the validity of the findings, they served to inform the decisions made for the main study. In the main study, the genre and prompt were kept the same for a within-groups design because this reduced the effect that different topics would have on the amount of target vocabulary use. The writing that was analyzed was completed under testing conditions to control time on task and to ensure that the students did not refer to vocabulary lists or dictionaries.

Pilot Study 2

Pilot Study 2 examined the changes in students' L2 vocabulary-learning beliefs and strategies (Wong, 2017). In this case study of six participants I examined their vocabulary learning beliefs and strategies for one 15-week academic semester. The participants were enrolled in a class in which Quizlet (2017) was a provided resource for vocabulary self-study. Each student was assigned several words and two tasks to: (a) find the meanings of the words, and (b) construct original sentences to aid understanding and show appropriate use. These tasks resulted in a student-generated Quizlet set that was intended for use in test preparation.

Pilot Study 2 Research Questions. Pilot Study 2 addressed the following research questions:

1. What are students' beliefs regarding vocabulary learning?
2. What strategies do the students use to study vocabulary?
3. What can change/s in strategy use be attributed to?

Pilot Study 2 Method. The participants fell into two groups, a larger group (N = 43) that completed the survey (see below), and a smaller group (n = 6) that agreed to be interviewed at the completion of the term. The participants in Pilot Study 2 were first-year Japanese students and were from the same institution as the learners in the main study. The participants (16 male and 27 female students) were from two intact classes I taught and who took a vocabulary course in the fall semester. In the first class meeting, informed consent forms allowing their survey data to be used in this study were completed and returned by all the participants.

Interviewees. The six students interviewed were recruited from the survey participants. The interviewees were six first-year female students who were informed of the nature of the study, the overall topic of the interview, and their rights as participants. The six students signed consent forms (see Appendix D for the informed consent form used for Pilot Study 2). I met with the participants individually prior to the start of the study in order to explain the purpose of the study and the terms stated in the informed consent form. The six interviewees were given the opportunity to seek clarification and to ask questions regarding the study before giving consent. The participants were selected based on a combination of high motivation and in-class participation, increases in their Vocabulary Levels Test (VLT) or GTEC scores, their willingness to be interviewed, and their availability at the end of the academic year. At the time of the interviews, I had taught four of the participants for two semesters and the remaining two participants for one semester.

Instruments. The main instruments were a survey and follow-up interviews of the six participants. Other instruments (GTEC, TOEIC, VLT) were used to support the findings of the survey and interviews (see Wong, 2017, for details).

The survey required participants to provide information regarding their learning histories, and vocabulary learning beliefs and strategies (see Appendices E and F for the English and Japanese versions of the language learning and vocabulary survey, respectively). It was administered twice to see if there were changes in participants' vocabulary learning beliefs and strategies prior to the start of the course and again at the completion of the course. Any changes relating to vocabulary learning beliefs and strategies, as well as comments regarding those changes, were noted for additional probing in the interviews.

Semi-structured interviews based on the participants' pre- and post-surveys were conducted at the end of the course to gain an understanding of the learners' perceptions of vocabulary study and the strategies they used (see Appendix G for Pilot Study 2's semi-structured interview questions). The structure consisted of questions about their English language histories, overseas experience, how they studied vocabulary in their secondary school education, their beliefs as to what types of vocabulary should be studied in English class, and the strategies they used to study vocabulary. In addition, the interviews incorporated areas from the surveys where the participants changed their responses. For example, a change in opinion regarding what students should study in English class was identified for further probing in the interview (see Appendix H for a sample of the interview question sheet incorporating reported changes used in Pilot Study 2). Interviews ranged from 30 to 45 minutes.

Pilot Study 2 Data Analysis. The interviews were audio-recorded and transcribed in their entirety for analysis. Because of my limited Japanese ability, interview questions were asked in English, but the participants were encouraged to answer in either Japanese or English as fully as possible. When the students answered in Japanese, my understanding of the answer was checked and confirmed by an L1 Japanese colleague. The themes outlined in the research questions (i.e., beliefs and strategies) were identified. The interview transcripts were reviewed multiple times to facilitate both deductive and inductive analyses. Through the deductive approach, the predetermined themes—vocabulary learning strategies, beliefs, and reported changes—helped identify relevant excerpts from the transcripts. The inductive analysis allows other themes that were not previous identified to be recognized.

Following the deductive analysis, the coded interview transcripts and survey summaries were analyzed for additional themes. Commonalities among the participants' interviews as well as explanations regarding changes in vocabulary strategies and beliefs, or academic behavior was identified. See Appendix I for a sample of a coded survey summary and interview.

Pilot Study 2 Results. The follow-up interviews revealed that use of the Quizlet (2017) set was irregular and the Quizlet app was not favorably viewed by the participants. Three of the participants in Pilot Study 2 reported that, despite the availability of the Quizlet set, they elected to use other means of studying vocabulary. Their reasons included a preference for writing on paper (two participants) and a desire for more control regarding the information for each Quizlet entry. The latter is a major criticism because self-generated study materials can be argued to be more effective than materials provided by a third-party, as posited by the Involvement Load Hypothesis (Hulstijn & Laufer, 2001). Research in word card use has also discussed this particular conundrum: Nation and Webb (2011) have asked if it is more effective for language learners to spend time creating their own word cards, or to spend that time on more study repetitions. At best, Quizlet was reported by the participants to be used mainly as a cramming tool rather than as a resource for regular vocabulary study, which, as pointed out in Chapter 2, is the least effective learner strategy for long-term retention (Dempster, 1987; Kornell, 2009; Pimsleur, 1967).

The following themes also emerged (inductive) from the interviews (see Appendix J for interview excerpts arranged by inductive themes):

- English will be useful primarily in their working lives and, to a lesser extent, their social lives.
- Speaking or communication are the most important part of language to learn and develop.
- All feel that they have vocabulary and fluency deficiencies.
- Outside of work, they are unclear as to how they can use English in Japan beyond "helping tourists".
- Exposure to English and perhaps some form of indirect studying is socially driven (i.e., Facebook, LINE, Instagram, Twitter) particularly among those who have overseas experience.
- Vocabulary study strategies or techniques are linked to the perceived purpose of learning those words (i.e., word lists of Quizlet for tests, and accidental learning through negotiation of meaning for socializing).

In summary, the interviews revealed that English is perceived to be of most use as a tool in the workplace with little relevance elsewhere. As a result of this, communicative competence was considered the most important skill to possess although all interviewees remarked that they were unsatisfied with their level of vocabulary knowledge and their ability to convey meaning. These comments were consistent among all interviewees despite high motivation and reasonably high proficiency. However, learners who had overseas experience in the form of study abroad, participating in exchange programs or homestays tended to use English more socially online. This social aspect of English use was attributed to greater English exposure that was not centered around language learning but, rather, for communicative purposes. It is interesting to note that the vocabulary study strategies employed reflected the purpose of learning those words. Learners who were concerned with English in the workplace used word cards or lists. However, learners who were using English in social contexts, while not deliberately studying vocabulary, learned word through negotiation of meaning.

Conclusions Drawn from Pilot Studies

The analysis of the writing samples in Pilot Study 1 revealed that the group that received reminders demonstrated a greater use of target vocabulary in their writing. This difference was significant compared to those in the non-reminder group. However, the study revealed that the implementation and administration of the reminders and essays require more care and consideration than was taken for Pilot Study 1. Therefore, the writing prompts for the Tracking Essays were controlled and each administration contained the same prompt in the main study. The Tracking Essays were conducted under exam conditions, thereby controlling time on task, dictionary use, and reference to vocabulary lists.

The results from Pilot Study 2 showed that students prefer the vocabulary learning strategy that they are used to. In most cases, this is the use of word lists or word cards. Even with these strategies, many students rely on cramming rather than the more effective spacing strategy. Therefore, the main study involved user metrics collected from the Quizlet (2017) website, self-reported data on Quizlet use, and the Moodle (2016) learning management system regarding time spent studying vocabulary using Quizlet and taking practice vocabulary tests on Moodle. Examining the amount of time each student spends on these two tasks can provide insights regarding the observed use of target vocabulary in the Tracking Essays and its relationship to the participants' vocabulary study strategy.

Together, the two pilot studies confirmed the feasibility of the study and provided preliminary insights into the transfer of vocabulary learned in one course to writing for another course. They also provided an opportunity to refine the data collection for this study (see Chapter 3, Methods). Next, the Gaps in the Literature section below will explain the rationale.

Gaps in the Literature

The above review of the literature provides an overview of current research related to learning in general and vocabulary learning in particular and indicates how these might transfer to written texts. Longstanding, established theories such as Level of Processing (LOP) (Craik & Lockhart, 1972), the Involvement Load Hypothesis (Hulstijn & Laufer, 2001), and Technique Feature Analysis (Nation & Webb, 2011) provide possible explanations as to why certain tasks might promote or encourage long-term vocabulary retention. Using tasks that make use of these frameworks, coupled with effective study techniques such as spaced learning and pushing students to use the target language, students are able to optimize their approach to learning vocabulary. Upon successful vocabulary learning, the next challenge is to transition from receptive to productive use, specifically free active use (Laufer & Paribakht, 1998). The ability to use words voluntarily is perhaps the ultimate goal of vocabulary learning.

Research in L2 writing suggests that vocabulary is a key aspect in the evaluation of writing quality (e.g., Astika, 1993; Santos, 1988). As such, university learners should focus their time and effort on studying vocabulary that is relevant to their field of study. Additionally, it is important that learners adopt a study method that includes seeing words in a variety of contexts, such as by using concordancers (e.g., Chan & Liou, 2005; Kaur

& Hegelheimer, 2005; Sun, 2007) and practice good writing habits by applying self-regulation techniques (Byrnes & Manchón, 2014).

The first gap concerns salience of treatments in experimental studies. I have not located any SLA studies that discuss the degree to which the treatments were noticed or understood by the participants, particularly for treatments that might easily be overlooked. If the participants in a study do not notice or understand a treatment, any observed or measured effects in the study cannot be attributed to the treatment. Thus, it is important to get a better picture of how salient a treatment is in order to minimize the amount of uncertainty when interpreting results.

The second gap concerns time spent studying vocabulary using particular study methods leads to the ability to freely produce language. Previous studies focusing on study time have addressed this issue from several perspectives; the regularity of study (i.e., crammed/massed/spaced learning), the use of word cards or words lists (e.g., Nakata, 2011; N. Schmitt, 1997; Wissman, Rawson, & Pyc, 2012), the effectiveness and efficiency of using word cards and word lists (e.g., Fitzpatrick, Al-Qarni, & Meara, 2008; Laufer, 2003b; Nation, 2001; Webb, 2009), and the retention of vocabulary learned using word cards (e.g., H. Bahrick et al., 1993; H. Bahrick & Phelps, 1987). Despite the large body of research in this area, there remains a gap concerning whether time spent studying vocabulary using particular methods have a direct bearing on the learners' ability to freely produce language.

The third gap concerns the transition from receptive to productive vocabulary use, specifically, vocabulary learning using word cards to the use of those words in writing assignments. Much of the literature has examined spoken output rather than written output (e.g., Hilton, 2008). Where other researchers have investigated written output, the participants produced a narrow range of vocabulary (Laufer, 1998) or the essay topics were cognitively challenging (Should governments be allowed to limit the number of children in families? Can technology replace traditional face-to-face teaching in the classroom?) that potentially affected the vocabulary the learners were able to use due to a higher perceived risk and/or lower confidence (Coxhead, 2008). A narrow range of produced vocabulary would make any inferences regarding the transition from receptive to productive use more difficult. Additionally, in those studies, the writings were analyzed using the Lexical Frequency Profile (LFP), which indicates which words in the samples come from different frequency levels. This does not accurately reflect whether the writing samples contain words that were previously unknown.

The fourth gap concerns transfer of knowledge, the ability to apply what is learned in one situation to another. In the context of this study, transfer of knowledge is defined as the transfer of vocabulary knowledge learned in one course to productive use in a writing task in a second course. Transfer has been assumed but has not been confirmed despite the above research in vocabulary study and use. In the case of vocabulary learning, educators assume that if learners successfully study vocabulary, with success typically being determined by a vocabulary test, then that vocabulary will be applied in other contexts. For this study, this transfer presumes that the target vocabulary is used correctly. However, I have located no research investigating whether vocabulary learning in one course transfers to writing done in another course. Researchers that have investigated how vocabulary learning affects writing taught both skills in the same course, which might have led the instructor to influence the outcomes by inadvertently alerting the participants to the true purposes of the study (e.g. Laufer & Paribakht, 1998; Miralpeix and Muñoz, 2018; Stæhr, 2008).

The fifth gap concerns learner beliefs regarding the usefulness of L2 study. While motivation has been extensively researched, I have not located any studies that have discussed how perceived usefulness affected the interpretation of the results. It is important to understand how perceived usefulness influences the success of vocabulary learning treatments. This comparison has been an under researched area of L2 research.

Purposes of the Study

The first purpose was to evaluate the salience of the vocabulary use reminders in the three treatment conditions. The reminders were presented orally and visually using a poster to all participants in the treatment conditions (R, R+, R++). The R+ condition completed a task requiring them to select vocabulary they thought would be useful in writing assignments. The R++ condition completed the same task as the R+ condition, but also had to write the chosen words in a space provided on worksheets for their writing assignments. Salience of the reminders was ascertained through the interviews conducted at the end of the first and second semesters. Additionally, the amount of vocabulary study time reported by the participants and collected from Moodle and Quizlet were analyzed to evaluate if there were any significant differences between the conditions.

The second purpose was to evaluate the effect that study time had on vocabulary learning and retention. This comparison was made using a mixed ANOVA to analyze the participants' receptive and productive vocabulary ability estimates (Rasch logits) with supporting data coming from correlation analyses of the amount of target vocabulary used in their tracking essays and their Vocabulary Inventory Surveys.

The third purpose was to investigate the effects of the treatments on the developmental trajectories of language proficiency and writing. Vocabulary knowledge has been shown to be a strong predictor of language proficiency (Miralpeix & Muñoz,

2018) and writing scores (Astika, 1993; Jordan, 1997; Santos, 1988). Therefore, it was necessary to see which treatment provided optimal results.

The fourth purpose was to assess the usage of the target vocabulary in the writing task. I have not located any studies that have examined the degree to which words are used correctly in L2 writing in addition to amount of target vocabulary usage. Hence, a chi-square test for independence was carried out.

The fifth purpose was to evaluate the perceived usefulness of the vocabulary study and its potential influence on the results. Language learners do not always understand the intended purposes of vocabulary study and, as a result, their performance suffers. The relationship between motivation and L2 learning beliefs is still not fully understood. Therefore, interviews were conducted to reveal whether the vocabulary study was perceived to be useful with regard to their writing.

Hypotheses

Five hypotheses arise from the review of the literature. First, the reminders will be more salient in the stronger treatments due to the inclusion of tasks designed to facilitate noticing the target vocabulary and guidance towards their use in writing. Swain (1998) argues that by noticing the gaps in their linguistic ability, learners can make the conscious decision to address these deficiencies. By pushing the learners to use the target vocabulary that they have chosen in their essays, this is encouraging generative use of those lexical items (Meyers, 2013). Furthermore, the Involvement Load Hypothesis (Hulstijn & Laufer, 2001) and Technique Feature Analysis (Nation & Webb, 2011) of the tasks suggest that this will result in more time being spent on studying the target vocabulary using Quizlet and Moodle.

The second hypothesis is that increased vocabulary study time will improve vocabulary learning and retention for the reminder groups. Evidence of learning and retention will be obtained from the results of the New General Service List Test–Receptive (NGSLT-Receptive) and the New General Service List Test–Productive (NGSLT-Productive). This hypothesis is supported by studies examining frequency effects in language acquisition (see Chapter 2) and the spacing effect (e.g., N. Ellis, 1995; Nakata, 2015). Subsequently, this improved vocabulary knowledge will aid the transition from receptive to productive knowledge (Henriksen, 1999; Hilton, 2008; Laufer & Paribakht, 1998). Vocabulary study time is expected to correlate positively with the vocabulary test scores and the amount of target vocabulary use in the tracking essays.

The third hypothesis is that improved vocabulary learning and retention for the reminder groups will lead to better GTEC and tracking essay scores. Studies have shown strong, positive correlations between vocabulary and reading comprehension (Coady, Magoto, Hubbard, Graney, & Mokhtari, 1993; Koda, 1989; Laufer, 1997). Similarly, several researchers have shown that vocabulary is a good predictor of writing scores (Astika, 1993; Jordan, 1997; Santos, 1988). As a result of improved vocabulary learning and retention, the participants' vocabulary test scores (NGSLT-Receptive and NGSLT-Productive) are expected to correlate positively with their GTEC and FACETs fair averages of the tracking essays.

The fourth hypothesis is that improved vocabulary learning and retention for the reminder groups will enable the participants to use the target vocabulary more accurately in the tracking essays. The task in the Remind + Notice (R) and Remind + Use (R++) conditions required students to choose useful vocabulary for their essays. Vocabulary has been established as one of the main predictors of writing scores (Astika, 1993; Jordan, 1997; Santos, 1988). Consequently, participants in stronger treatments will demonstrate more correct usage of the target vocabulary. The higher incidence of correct vocabulary use will lead to better overall writing scores (JACET writing rubric, lexical complexity, standard length words). Vocabulary study time was expected to correlate positively with amount of correct target vocabulary use.

The fifth hypothesis is that as a result of better vocabulary learning and retention, higher vocabulary, English proficiency, and writing scores, learners who were reminded will report greater perceived usefulness of the vocabulary study as it relates to their writing than those with no reminder.

CHAPTER 3

METHODS

The purposes of this chapter are to introduce the context of the study, including the setting, provide details about the courses from which participants were drawn, provide information about study participants, outline the methodological framework, to present the instruments, and to explain the rationale behind the approaches used to analyze the results. In this section, I first detail the setting and context of the site and how these were incorporated into the study design (see Research Design below).

Context for the Study

The Japanese University Ranking System (*Hensachi*)

The *hensachi* is a score assigned to individual students or school departments based on student performance in a national test standardized to the national mean (McLean, Hogg, & Kramer, 2014). A *hensachi* of 50, therefore, represents the national mean and a student or university with this ranking would be at precisely at the center of its group. A score one standard deviation over the mean is 60, and one standard deviation under the mean is 40. *Hensachi* scores can range from 20 to 80 and, for universities, 94.5% of all university departments fall between 30 and 70 (Newfields, 2006). To minimize confusion for those unfamiliar with the *hensachi* rank, it has been suggested that a better translation might be *standardize rank score* (Newfields, 2006). The *hensachi* is used by high schools as a way to recommend institutions to their students.

Setting and Courses

The study took place in a management department in a private university in western Japan. The ranking for this department at the time of this study, which is known in Japanese as *hensachi*, was 58 (Benesse, n.d.), roughly equating to the 85th percentile (Newfields, 2006).

Compulsory English Courses. All participants took three compulsory English courses, a vocabulary course, a content course, and a speaking course. The vocabulary course and content course were included in this study. Each course met for two semesters. Each class period was 90-minutes each week, and each semester was 15 weeks. The courses were independent, meaning the content of one course was not linked to or integrated with others at the curricular level. Although rare, there are cases when an instructor teaches the same students in more than one of the three courses in a single semester. Each semester, the vocabulary and content courses were streamed by English language proficiency. In the speaking-focused course, the students were assigned to classes based on their student numbers. This unstreamed course was not included in this study because the coursework did not include vocabulary development or writing instruction and, thus, had no bearing on the discussion or outcomes. The vocabulary course and the content course, the two courses, that provided the setting for the study, are described next.

Vocabulary course. Each semester, all students were streamed by ability into one of eight classes that were focused on study skills, with particular attention paid to vocabulary development. The vocabulary classes met once a week over each 15-week semester, totaling 22.5 hours of classroom time each semester.

For this study, the students were streamed in vocabulary course 1 based on scores on the Global Test of English Communication (GTEC) (Benesse, 2004) and a modified version of the New General Service List Test-Receptive (NGSLT-Receptive) (Stoeckel & Bennet, 2015) scores (see Participants and Instruments sections below).

In addition to the NGSLT-Receptive, the test used by the institution for class placement, the students were given the New General Service List-Productive (NGSLT-Productive), which was developed and administered for this study (see Instruments below). Together, the NGSLT-Receptive and the NGSLT-Productive served as diagnostic tests to assess strength of knowledge of the vocabulary for this study.

Table 9 shows the scores used to place students into the vocabulary course based on their GTEC and NGSLT-Receptive scores. In the second semester, they were streamed into vocabulary course 2 based on their TOEIC scores and in-class performance in the first semester. The classes were divided into two blocks of four classes each; classes 1 through 4, and 5 through 8. Students with the highest scores were placed in classes 1 and 5, and those with lower scores in classes 4 and 8 (see Participants below).

Table 9*. Initial Placement of Participants by GTEC and NGSLT-Receptive Scores*

Block	Class	GTEC	NGSLT-Receptive
A	1	214-303	32-50
	2	193-238	34-48
	3	172-207	33-47
	4	121-213	22-46
B	5	214-284	40-49
	6	192-242	36-46
	7	159-216	29-45
	8	105-189	23-39

Note. GTEC = Global Test of English Communication. NGSLT-Receptive = New General Service List Test-Receptive, a 50-item version of the New General Service List Test.

The blocks consisted of an approximately equal number of students. Each block of classes met at the same time with the schedule determined by the university administration office. During the second semester, students were moved into a higher or lower class based on their TOEIC scores. (See the Participants section below for a discussion of the distribution of the students.)

Vocabulary self-study, self-test, and testing. In vocabulary course 1, the students were expected to enter knowing the first 500 words of the NGSL. During the course, they were expected to study the second 500 words (NGSL 501-1000). In vocabulary course 2, they studied the next 500 words (NGSL 1001-1500).

Each term, the list of 500 words was divided into 10 sublists of 50 words, with students studying one sublist each week. The vocabulary tests (see Instruments below) assessed the sublists studied in the previous three or four weeks and were scheduled in weeks 6, 10, and 14 each semester. Table 10 displays when the sublists were studied and the testing schedule during this project.

For each sublist of 50 words, three online resources were prepared: definition lists, Quizlet (2017) sets (see Appendix K for a sample of a Quizlet Vocabulary Set), and Moodle (2016) vocabulary quizzes. The definition lists contained the target words, followed by their parts of speech, English definitions, and Japanese translations. The Quizlet sets showed the target words with their corresponding English definitions. The primary study activity used by students in Quizlet was the flashcard utility. Quizlet displayed the English definition of a target word and, when the user tapped the screen, it displayed the target word before proceeding to the next definition.

Table 10. *NGSL Vocabulary Self-study, Moodle Self-test Quizzes, and Vocabulary Testing Institutional Schedule*

Lesson	Semester 1			Semester 2		
	NGSL list studied	Moodle quiz	Vocabulary test	NGSL list studied	Moodle quiz	Vocabulary test
1						
2	(501-550)	1.1		(1001-1050)	2.1	
3	(551-600)	1.2		(1050-1100)	2.2	
4	(601-650)	1.3		(1101-1150)	2.3	
5	(651-700)	1.4		(1151-1200)	2.4	
6			Test 1 (501-700)			Test 4 (1001-1200)
7	(701-750)	1.5		(1201-1250)	2.5	
8	(751-800)	1.6		(1251-1300)	2.6	
9	(801-850)	1.7		(1301-1350)	2.7	
10			Test 2 (701-850)			Test 5 (1201-1350)
11	(851-900)	1.8		(1351-1400)	2.8	
12	(901-950)	1.9		(1401-1450)	2.9	
13	(951-1000)	1.1		(1451-1500)	2.1	
14			Test 3 (851-1000)			Test 6 (1351-1500)
15						

Note. NGSL = New General Service List. Numbers in parentheses indicate range of NGSL vocabulary.

Course developers obtained the definitions from the downloadable NGSL definition set. Using the definitions provided by Browne et al. (2013b) ensured that the definitions associated with the target vocabulary matched their stated frequency. These two resources were prepared to support the participants' vocabulary learning and comprised a weekly self-study task. All Quizlet (2017) sublist sets were released to students on the first day of the course. All students in the program were told to study a particular sublist in their vocabulary course each Friday (see Table 10).

While the use of the definition lists and Quizlet sets were optional in their weekly self-study, the students in each class were required, as part of their course grade, to complete the online Moodle vocabulary quiz that corresponded to each 50-word list. These quizzes, along with the definition lists and links to the Quizlet sets, were made available on the university's Moodle learning management system, which was accessible to all students. The quizzes were in a multiple-choice format, which required the students to complete a sentence by choosing the correct word from three options. The sentences were written by course instructors. Care was taken during the development of these quizzes to ensure that the meaning of the target word matched the definitions and translations on the definition list. Figure 1 shows an example item from a Moodle practice quiz.

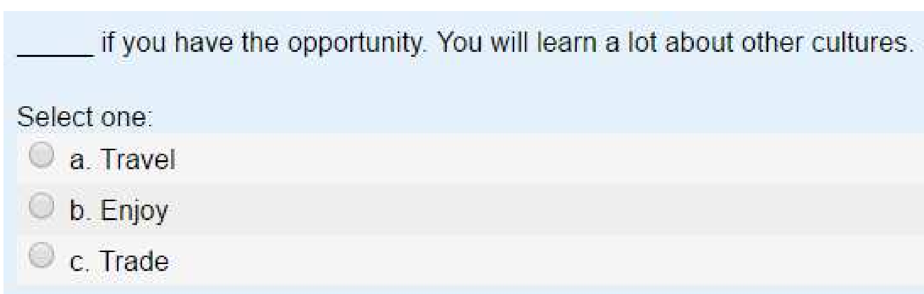

Figure 1. Vocabulary quiz item as it appears on a Moodle (2016) practice quiz.

The on-line Moodle (2016) vocabulary quizzes were made accessible at noon on the day the students had the vocabulary course (Friday). This timing allowed the students to study the words introduced that day in class and take the Moodle practice quizzes any time before the following week's class meeting, giving all students one week to complete the quiz. The students could take the quizzes multiple times, which enabled them to take the quiz as many times as they wished before the next class meeting. As the quizzes were out-of-class tasks, which made it impossible to eliminate copying or the use of study materials, quiz scores were not included in the students' grades. However, a reasonable attempt at taking the quiz, which was determined by time on task as reported by Moodle, was required.

During Weeks 6, 10, and 14 (see Table 10), the students complete in-class, paper-based vocabulary tests to demonstrate their command of the target vocabulary. The tests consisted of three sections: definition-matching, sentence completion, and writing original sentences (see Appendix C for examples of the test sections). For the construction of the vocabulary tests, the following test specifications were used by course developers. First, from the 150 words for that test, 35 words were randomly selected to appear as key items. The remaining 115 words were used as distractors. For the 35 key items, 15 were randomly assigned to the matching section, and 10 each to the remaining multiple-choice, and sentence-writing test sections. For the matching section, the 15 items were divided into five groups. For each of these groups, three distractor items were selected. Definitions were copied from those used on the NGSL list. Sentences were generated by the course instructors for the multiple-choice section. Care was taken to ensure that the part of speech of the key item was the same as in the NGSL list and that the words contained in the sentence were at the same frequency level or higher than that of the key item. For the sentence-writing section, the participants were instructed to write a sentence demonstrating that they knew the meaning of the key item as well as how to use it appropriately. The participants were given 20 minutes to complete the vocabulary test and use of notes or dictionaries were not allowed.

In summary, following course guidelines, the students were instructed to study 50 words per week. They could use the prepared definition lists and/or Quizlet sets to study. Before the next class meeting, they had to take a vocabulary quiz that was available on Moodle. After studying three or four 50-word lists, they were given a vocabulary test in class under test conditions. The test was graded by the course instructor. Table 10 shows the study and test schedule for each semester.

Content course. The students were instructed in several genres of writing in the content course. For each genre, the content provided the background information necessary to complete the writing assignments. Similar to the vocabulary courses, the students were streamed by ability, although these criteria differed slightly. The classes met twice a week for 15 weeks, totaling 45 hours of classroom time each semester. In content course 1, taken in the first semester, the students were streamed by their GTEC and Tracking Essay 1 (see Instruments below) scores. In content course 2, taken in the second semester, they were streamed by their TOEIC and Tracking Essay 2 scores.

The first semester content course had 10 classes, as opposed to eight classes for the other two compulsory courses, due to the presence of students enrolled in a study abroad program that also had to take the same course, which resulted in additional students in content course 1 only. These students were distributed across the 10 sections of the content course, with all students streamed into five levels based on ability.

The study abroad students did not take the content courses in the second semester, which had eight streamed classes, matching the number of vocabulary courses offered. In summary, while all the courses were streamed by ability, content course 1 was divided into five levels with two classes each due to the addition of students in the study abroad program. The vocabulary and speaking courses consisted of four English proficiency levels with two classes each.

Content course: Writing genre, assignments, and instruction. In content course 1, the students were instructed in the procedure, report, and narrative genres of writing. By the end of the semester, each student completed one writing assignment for each genre. Content course 2 revisited the report and narrative genres covered in the first semester and added the recount genre for a total of three writing assignments throughout the semester. Table 11 shows the writing assignments by semester.

Table 11. *Genre Writing Assignments in the Content Course for Semesters 1 and 2*

Semester	1	Week Taught	2	Week Taught
Writing Genres	Procedure	4	Recount	2
	Report	7	Narrative	5
	Narrative	10	Report	10

As part of the content courses, the students were instructed in the following genres of writing: procedures, reports, narratives, and recounts. Procedures give instructions for accomplishing a particular task. Common examples of procedures are recipes, instructions for assembling furniture, and instructions for wearing traditional clothing. Reports are used to provide information to readers by giving facts. This particular genre was selected for inclusion in the curriculum as the content courses have a strong research focus and knowing how to write a report was deemed essential. Narratives, unlike the other genres, primarily entertain and engage readers in an imaginary experience. Narratives can take the form of short stories, picture story books, fairytales, myths, and science fiction stories. The last genre, recounts, retells past events or experiences. A key feature of recounts is that events or experiences are written in the order they happened.

For each of the genres, the lessons followed the same series of steps. First, the students were given a model of the genre for general comprehension. Second, the teacher guided them in deconstructing the written model to examine the text and determine how it was produced. Third, the teacher led a joint construction exercise where the framework and target language were first elicited, followed by the students using a worksheet containing key points to collectively generate the full sentences and paragraphs. Finally, they worked independently to produce a first draft of a text, received feedback from their teacher and classmates, revised the draft, and then submitted a final draft. A set of the teaching materials used for the recount genre is provided in Appendix L.

Speaking courses. The speaking-focused courses consisted of class groups that were unstreamed by language proficiency, with classes decided by student number. The speaking course classes met twice a week for 15 weeks, totaling 45 hours of classroom time each semester. In both semesters, these courses involved spoken production, both monologue and discussions, as well as emphasis on formal presentations. As stated above, the speaking courses were not included as these do not involve vocabulary study or writing development.

Participants

The participants (N = 178) were drawn from the first-year students in the compulsory English program in the management department. The participants were 18 or 19 at the start of study. All participants were native speakers of Japanese and entered the university through entrance exams, high school recommendations with oral interviews, or were in the university school system since elementary school due to the so-called elevator system that allows students to automatically move to the next educational level. All participants had at least six years of formal education in English in Japanese junior and senior high schools combined (see Chapter 2 for a discussion of Japan's education system). However, their backgrounds and language learning histories varied, with the general English proficiency level at the time of entry into the university ranging from 105 to 303 on the Global Test of English Communication (GTEC) (Benesse, 2004), with a mean score of 204.00 and standard deviation of 35.08, which translates roughly to Test of English for International Communication (TOEIC) (ETS, 2017) scores ranging from 147 to 714, with a mean score of 430 based on the on-line conversion table provided by Benesse, 2017.

The participants were provided with an informed consent form on the first day of class (see Appendices M and N for English and Japanese versions of the informed consent form, respectively), before they completed the placement tests that determined their course placement. The informed consent form provided information about the general purposes of the study and let the participants know that the data they provided would be used for research purposes. The form also allowed them to opt out of the study. While the study abroad program students also completed the informed consent forms and received the same placement tests, instructions, materials, and reminders, these students were excluded from the data analysis. Figure 2 shows the number of participants, minus the 22 from the study abroad course, in the vocabulary course and content course in the first semester. All students in all courses agreed to participate.

Interviewees

Twenty participants were interviewed once after each semester for a total of two interviews each. Interviewees were selected based on the recommendation of their content course instructors. All instructors were asked to recommend two candidates from each of the classes they taught based on the students' in-class performance and the instructors' perception of the students' work ethic and academic conscientiousness. The decision was deliberately made to solicit students that were engaged in their learning in order to determine the effects the degree of reminder had on students who were more likely to follow the teachers' instructions (i.e., a best-case scenario).

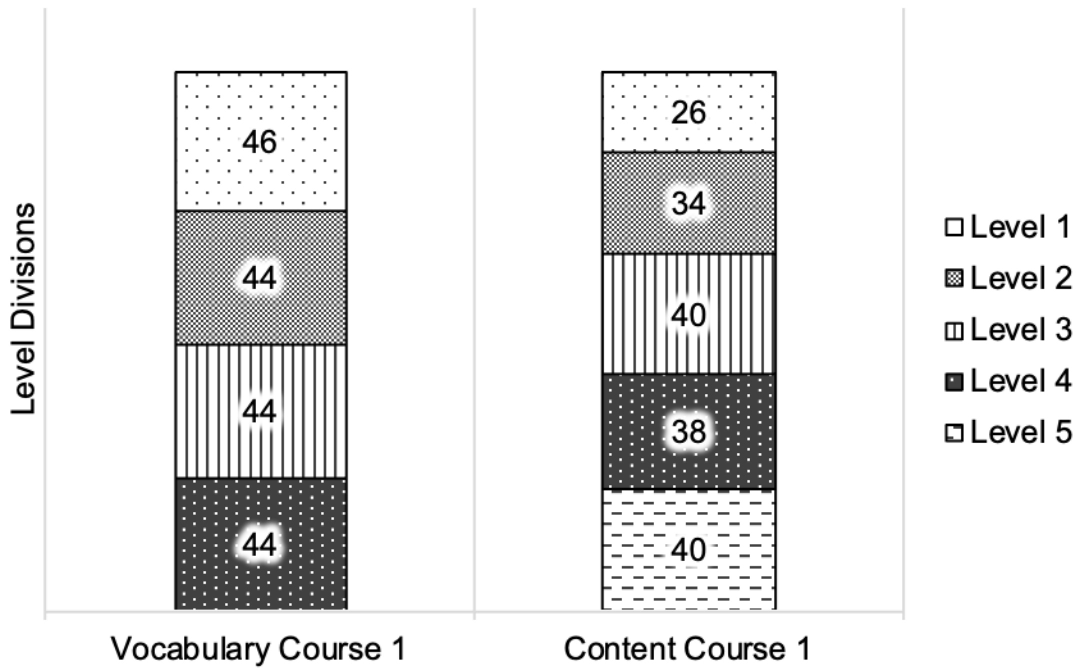

Figure 2. Number of students in each level in the vocabulary and content course in the first semester. Content Course 1 consisted of five levels due to an additional 22 students from a study abroad program. These students were not included in the study.

Participation in the interviews was voluntary and the interviewees were informed that no part of the study would affect their grades. They were asked to read and sign a more detailed informed consent form than that used with the main study participants (see Appendices O and P for the Japanese and English versions of the informed consent form that were provided to the interviewees). Interviewees were also informed of their right to withdraw from the study at any time and given my contact details to use if they had any questions. Students were informed of their right to access the data collected about them. To avoid biasing the results, the participants were not informed in detail about the purpose of the study before it was completed. Interviewees received a small remuneration for their participation (¥1,000 minus taxes, approximately $10 USD), in line with the institution's policy regarding student participation in research and other activities.

This process resulted in two participants from each class being selected for interviews (see Table 13, Study Conditions of the Participants). Four participants were selected from each of the three treatment groups, totaling 12, and eight participants were selected from the control group. All 20 participants completed each of the interviews.

Instruments

This study made use of instruments already in use in the department —including the NGSLT-Receptive, GTEC, and study materials— and those developed for the study within the program's parameters —the NGSLT-Productive, Tracking Essay, vocabulary self-study log, vocabulary inventory survey. Therefore, the instruments served three purposes: (a) to track the students' English language development throughout the program, (b) to place students into most, but not all, of their compulsory English courses, and (c) to provide the data for this study. All instruments developed for this study were constrained by institutional requirements, although some accommodation to the courses and testing procedures were made to facilitate data collection.

Tracking Essays

The tracking essays were required of all the enrolled students and were administered under test conditions before classes began in April, at the end of the first semester in July, and at the end of the second semester in January. The students were required to write each Tracking Essay in 30 minutes using the recount genre. No pre-writing tasks were used and the writing was done by hand.

The writing prompt provided in Figure 3 was used at the specified times during the study. The Deputy Director of the program, when approving this study, specifically indicated that the writing prompt could not explicitly direct the students to write about the same event for the second and third tracking essays. However, he also indicated that it was acceptable for students to decide on their own to write about the same event.

Sample # 1234567 Name: SAMPLE Name
Title: A memorable event in your life
Directions: Write about a memorable event in your life. Describe who you were with, what occurred, and when and where it took place. Write about it in the order it happened and give as much information as possible. Do not use a dictionary. You have 30 minutes. Please use the NGSL words you have studied in this writing.

Figure 3. Writing prompt used in this study with all reminder condition participants. The final sentence, "Please use the NGSL words you have studied in this writing" did not appear for the non-reminder group or for Tracking Essay 1. NGSL = New General Service List.

For each Tracking Essay, the students received a sheet of paper prepared with their student numbers and the prompt. The rest of the paper was lined. The paper was A3-sized, slightly larger than two U.S. letter-size pages. Japanese instructions were printed below the English instructions (see Appendix Q for a copy of the Tracking Essay sheet with English and Japanese instructions).

Initial Placement Tests

Upon entering the university, all students were given a battery of placement tests. These tests included the Tracking Essay, explained above, and three additional tests described below; a shortened version of the New General Service List Test (NGSLT), a productive version of the NGSLT (NGSLT-Productive), and the Global Test of English Communication (GTEC). The tests were administered sequentially in the order stated above and the total test time was 90 minutes. Some adjustments were made to the NGSLT, explained below, to fit into the time constraints of initial placement.

New General Service List Test—Receptive

The original New General Service List Test (NGSLT) (Stoeckel & Bennet, 2015) is a 100-item test of written receptive knowledge of the New General Service List (NGSL) (Browne, Culligan & Phillips, 2013b). Each item provides the target word followed by a short, decontextualized sentence. The test covers the frequency range of the NGSL and is divided into five sections of 20 items each with each section representing one approximately 560-word frequency band. Testing across the entire frequency range of the NGSL is necessary as the purpose is to measure the total vocabulary size of the test-taker and, thus, needs to measure frequency levels beyond the test-taker's likely vocabulary size. However, while the test measures receptive vocabulary knowledge, it does not indicate how well the known words could be used in speaking or writing.

The test consists of a stem plus 4 answer choices in a multiple-choice format (see Figure 4). The item stem consists of the word followed by a simple, decontextualized sentence containing the word. The sentence does not define the word and indicates the part of speech of the word. The sentence also limits the meaning of the word where words may have homographs or different senses. However, the test does not measure the ability to distinguish homonyms and homographs. As a result, the test measures knowledge of single words. The distractors should fit the stem grammatically and be from around the same frequency level as the correct answer. The distractors should not share core elements of meaning with the correct answer. The test items are equally weighted so, multiplying the test taker's final score by 28 will give their estimated vocabulary size.

Example 1:

charge: They are the **charges**.

a. important things to think about

b. prices for a service

c. good things

d. reasons

Figure 4. Example test item on the receptive New General Service List Test.

Administration of the test is not limited by time as the test is a measure of knowledge not fluency. Typically, 30 minutes is sufficient for the 100 items. However, the time provided by the university to administer the three placement tests necessitated the creation of a smaller 50-item version of the test by the program instructors. This test was constructed by extracting the odd-numbered items from the NGSLT, which maintained the same coverage of the New General Service List and, therefore, the same range of difficulty. The test was paper-based and students were given 20 minutes to complete it. This test is designated as the NGSLT-Receptive in this study. It was administered in April, at the start of the first semester, in September, at the start of the second semester, and finally in January, at the end of the second semester.

New General Service List Test—Productive

The productive New General Service List Test, hereafter, referred to as the NGSLT-Productive, is a 30-item test of productive knowledge of the NGSL (Browne et al., 2013b). The test was constructed by extracting 30 of the same words from the NGSLT-Receptive, six from each of the five frequency bands. Each item consisted of the Japanese definition of the target word followed by a translation of the sentence in Japanese that included the word. The tested word was then given as a blank, with the second letter provided, in an English sentence. Though it is more common for test makers to provide the first letter of a word if only one letter is given, this hint, along with the Japanese definition, was felt to severely reduce the productive demands on the participants for this test. This differs from Laufer and Nation's (1999) Productive Vocabulary Levels Test, which provides up to the first five letters of a target word. An example test item with translation is shown in Figure 5 (see Appendix R for the full 30-item NGSLT-Productive). Prior to use in the study, the NGSLT-Productive was piloted with nine student volunteers to ensure that there was only one acceptable answer per question.

The purpose of the NGSLT-Productive was to allow inferences to be made regarding the strength of vocabulary knowledge of words that were shared between the NGSLT-Receptive and NGSLT-Productive. Participants' ability to answer the test items in the NGSLT-Productive constitutes controlled active vocabulary as defined by Laufer and Paribakht (1998). The 30 items were selected by extracting the even-numbered items from the 50 items included in the institution's NGSLT-Receptive version of the test and items numbered 5, 15, 25, 35, and 45. This test was administered before the receptive test to prevent the participants from being trained or exposed to the target vocabulary in the NGSLT-Receptive. The number of items on the NGSLT-Productive was set by institutional requirements.

Example 1:
学校
これは大きい学校だ。
This is a big _c__________.

Figure 5. Example test item on the New General Service List Test-Productive. 学校 (gakkou) = school. これは大きい学校だ。(kore ha okii gakkou da) = This is a big school.

Global Test of English Communication (GTEC)

The GTEC is a computer-based test of English proficiency and is designed to measure English for business environments (Benesse, 2004). The purpose of the GTEC at the university was to obtain another measure of the students' English language proficiency prior to starting classes, with the GTEC results and Tracking Essay 1 scores used to place students into content course 1 and the GTEC results and NGSLT-Receptive scores to place them into vocabulary course 1. Access to the raw item-level scores for the participants was restricted by Benesse.

The GTEC is not as widely used in Japan as other language proficiency tests such as the TOEIC or TOEFL. The version of the GTEC used in this study consisted of two sections designed to measure listening and reading. Test takers had a total of 40 minutes for the two sections and the maximum total score is 500 points.

The GTEC was a required part of the English program. The students took it prior to starting classes in April, and again at the end of the second semester in early January. Despite this weakness, as the students took the GTEC twice, it provided a measure of English proficiency and a way to detect changes in proficiency.

Quizlet Self-Study Metrics

Through a paid teacher account, Quizlet provided metrics regarding each participant's visit to the website or use of the smartphone application as well as how they performed in each of Quizlet's study modes. Quizlet reported how many times each study mode was started and/or completed (see Figure 6 for an example of the output of this information for one student). This information provided insight into the regularity and quality of the participants' vocabulary study, which in turn partially explained gains in test scores. As Quizlet did not provide time-use metrics, the participants were asked to keep a record of time spent using Quizlet and other vocabulary study methods, each week in their vocabulary course (see Appendix S for the Vocabulary Self-Study Log).

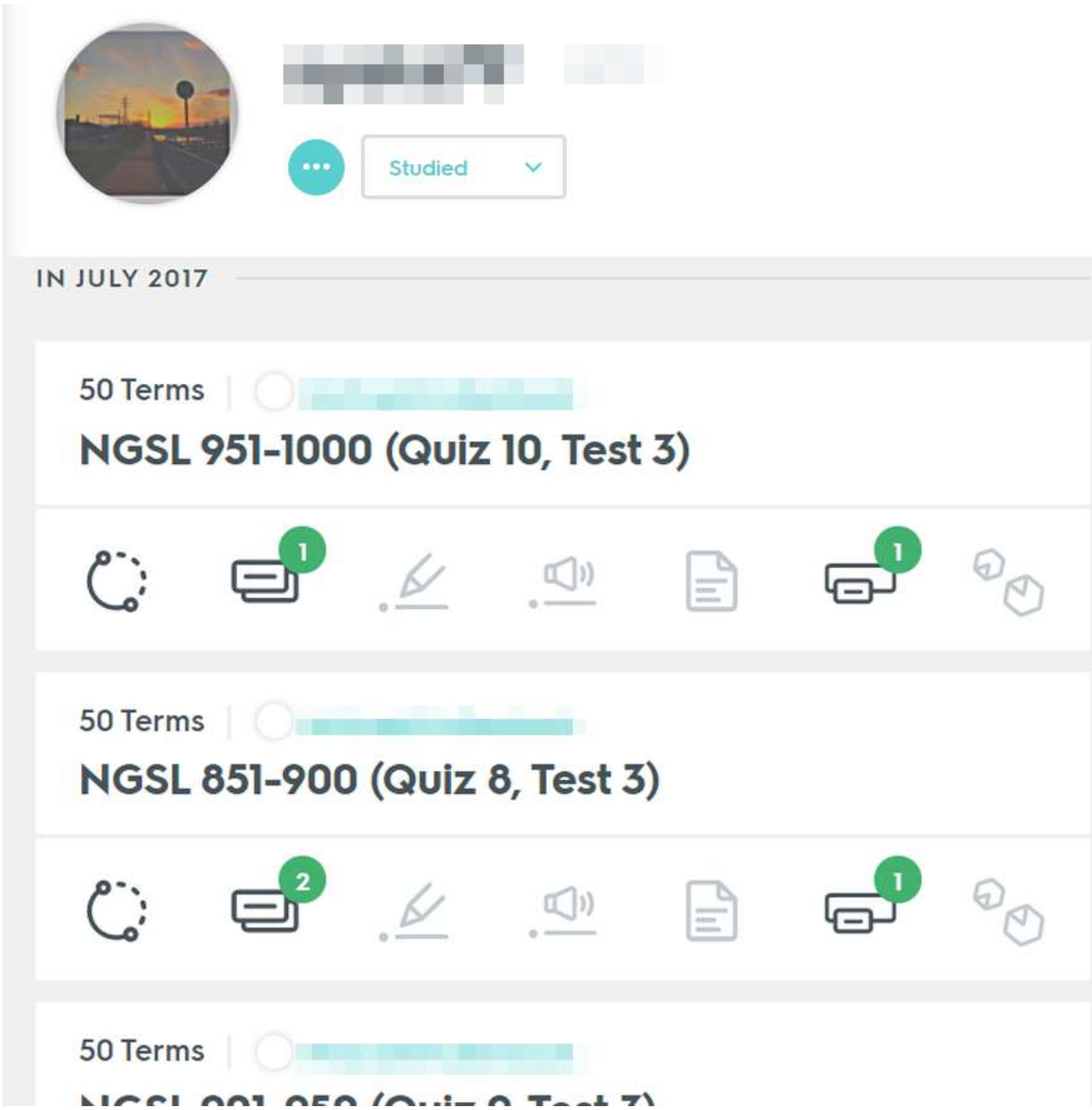

Figure 6. Quizlet (2017) page showing a student's use of the various study modes.

Moodle Self-Test Metrics

The Moodle (2016) learning management system was administered by the university and used by all instructors and students. The participants used Moodle to access the 10 weekly vocabulary quizzes as part of their vocabulary course. Moodle provides its own metrics regarding the time taken for the test, the number of questions answered correctly, and number of attempts made (see Figures 7 and 8 for examples of Moodle output for a group of students and for one student). This information, like the metrics obtained from Quizlet, was useful in assessing participants' commitment to their vocabulary learning. The settings for vocabulary quizzes were determined by the institution and these allowed the students multiple attempts during the one-week period when each quiz was available. In effect, the participants had an unlimited number of opportunities to attempt each quiz.

First name / Surname	Email address	State	Started on	Completed	Time taken	Grade/20.00	Q. 1 /1.00
[illegible] Review attempt	[illegible]	Finished	23 September 2016 12:35 PM	23 September 2016 12:40 PM	4 mins 53 secs	19.00	✓ 1.00
[illegible] Review attempt	[illegible]	Finished	23 September 2016 2:54 PM	23 September 2016 3:06 PM	12 mins 42 secs	17.00	✓ 1.00
[illegible] Review attempt	[illegible]	Finished	24 September 2016 10:34 AM	24 September 2016 10:38 AM	4 mins 8 secs	16.00	✗ 0.00
[illegible] Review attempt	[illegible]	Finished	24 September 2016 3:11 PM	24 September 2016 3:17 PM	6 mins 5 secs	14.00	✓ 1.00
[illegible] Review attempt	[illegible]	Finished	24 September 2016 4:19 PM	24 September 2016 4:25 PM	6 mins 21 secs	20.00	✓ 1.00
[illegible] Review attempt	[illegible]	Finished	26 September 2016 2:36 PM	26 September 2016 2:46 PM	9 mins 51 secs	17.00	✓ 1.00
[illegible] Review attempt	[illegible]	Finished	26 September 2016 2:50 PM	26 September 2016 2:55 PM	5 mins 32 secs	18.00	✓ 1.00

Figure 7. Moodle (2016) page showing seven students' general information, quiz status, date taken, duration, quiz scores, and the first of the individual item scores and whether answered correctly or not.

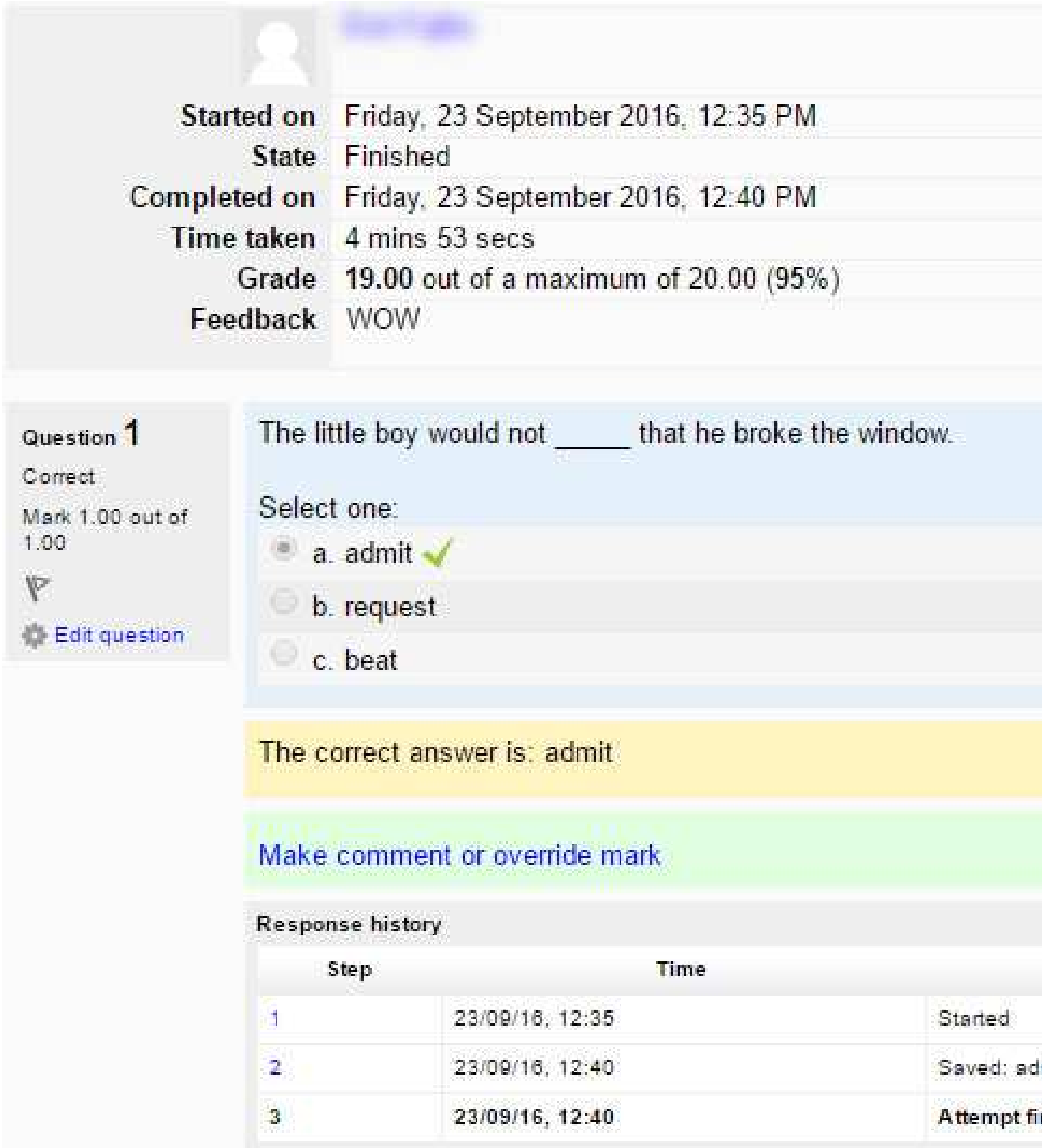

Figure 8. Moodle (2016) page showing part of a detailed report of a student's quiz.

Vocabulary Inventory Survey

At the beginning and end of the first semester and the end of the second semester, the students completed a vocabulary checklist of the second 500 words of the NGSL, the vocabulary covered in the first semester (see Appendix T for the Vocabulary Inventory Survey). They were instructed to place a check mark next to the words that they knew. In addition, they were told that if they took more than a few seconds to decide if they knew a word or not, they should not place a check mark next to that word. This measurement was to give a rough estimate of how much of the target vocabulary was known before the course began, how much their vocabulary knowledge changed over the semester, and how much was retained in the following semester.

Vocabulary Self-Study Log

At the beginning of each class meeting for the vocabulary course, the students were told to fill in a vocabulary self-study log to indicate how much time they spent studying the target vocabulary. They were asked to write the time spent studying the target vocabulary (a) the day before the class meeting, and (b) during the last week since the previous class meeting. They had to write these times for (a) study with the Quizlet smartphone application, and (b) other study methods such as checking the printed vocabulary list, using word cards, or practice writing the words or example sentences.

Figure 9 shows a sample of the self-study log. Students who were absent were not asked to supply missing data the following class.

Week 3	**Quizlet**	**Other**
How much time did you spend studying vocabulary *yesterday*?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary *last week*?	_____ mins.	_____ mins.

Figure 9. Example section of the vocabulary self-study log done at the beginning of each vocabulary course class meeting.

Interviews

Semi-structured interviews were conducted at the end of each semester to gain an understanding of the participants' learning experiences in the vocabulary and content courses and to gauge the saliency of the reminders (see Appendix U for the Semi-Structured Interview Questions). The structure consisted of questions about what they liked best and found most difficult about the courses, their vocabulary learning in vocabulary course 1 and how this affected their writing in content courses 1 and 2. The participants were also asked how frequently they studied vocabulary and to reflect upon their learning over the semester.

The interviews ranged from 21 minutes to 56 minutes. Interviews were conducted in English, but Japanese versions of the questions were provided in writing in order to ensure that they could be answered completely. The interviewees were told that they could answer in either English or Japanese. These data were used to answer hypothesis 1, the reminders will be more salient in the stronger treatments, and hypothesis 5, learners who were reminded will report greater perceived usefulness of the vocabulary study as it relates to their writing than those with no reminder. In the second interview, conducted after the end of the study, the interviewees in the R, R+, and R++ conditions were also asked about the effect of the reminder on their writing. This question was not asked in Interview 1, as it might have influenced the participants' behavior during the second semester of the study.

Technique Feature Analysis Scores of the Instruments

As described in Chapter 2, Technique Feature Analysis (TFA) can be used to examine the effectiveness of vocabulary activities (Nation & Webb, 2011). According to Nation and Webb (2011), a vocabulary teaching technique consists of five criteria categories: motivation, noticing, retrieval, generation, and retention. A technique can be provided with a score based on these five categories (see Table 1 in Chapter 2 for the full list of criteria). A highly motivating activity requires a clear vocabulary learning goal, provides a purpose to the learners for doing the task, and encourages learners to select the words themselves (autonomy). Furthermore, effective techniques focus the learners' attention on target words as well gaps in their own lexical knowledge (noticing). Learning of target vocabulary is aided when activities require the learner to retrieve the meaning or the form of words from memory and provide multiple opportunities for these retrievals (retrieval). Memory of target words can be further strengthened when the activity involves meeting or using the words in a new way (generation). Finally, effective vocabulary teaching techniques facilitate the linking of form and meaning, involves instantiation and imaging, and avoids interference (retention). Following these scoring criteria, the TFA score of each of the instruments is shown in Table 12 (see Appendix V for the detailed Technique Feature Analyses of all the instruments).

Research Design

The study was conducted longitudinally, and the participants were followed over the course of two semesters. A longitudinal design was selected because this allowed vocabulary learning to be reflected in the participants' writing. Figure 10 provides an overview of the project.

Table 12. *Technique Feature Analysis (TFA) Scores of the Instruments*

Instrument	Score	Points of note
Tracking Essays	7	Productive retrieval + recall + generative use
NGSLT-Receptive	4	Attention focused on target words
NGSLT-Productive	5	Productive retrieval
GTEC	N/A	
Quizlet Self-study	8	Motivates learning + multiple spaced retrievals
Moodle Self-study	5	Attention focused on target words + retrieval
Assignment Essays NR	2	
Assignment Essays R	4	Vocabulary learning goal + Awareness
Assignment Essays R+	5	Attention focused on target words
Assignment Essays R++	8	Generative use encouraged

Note. The TFA score for the GTEC is not applicable due to the format of the test (i.e., computer-based multiple-choice). NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive, GTEC = Global Test of English Communication, R = in class reminder condition, R+ = in class reminder plus locate words from most recent vocabulary list condition, R++ = in class reminder plus locate words from cumulative vocabulary list condition, NR = No Reminder group.

The study most closely fits the Concurrent Embedded Design (Creswell & Plano Clark, 2010) in which the participants' writing samples and other measures of language proficiency are carried out concurrently. The primary research instrument was the tracking essays (see Instruments below) collected in April and at the end of each semester (Tracking Essay 1, Tracking Essay 2, Tracking Essay 3).

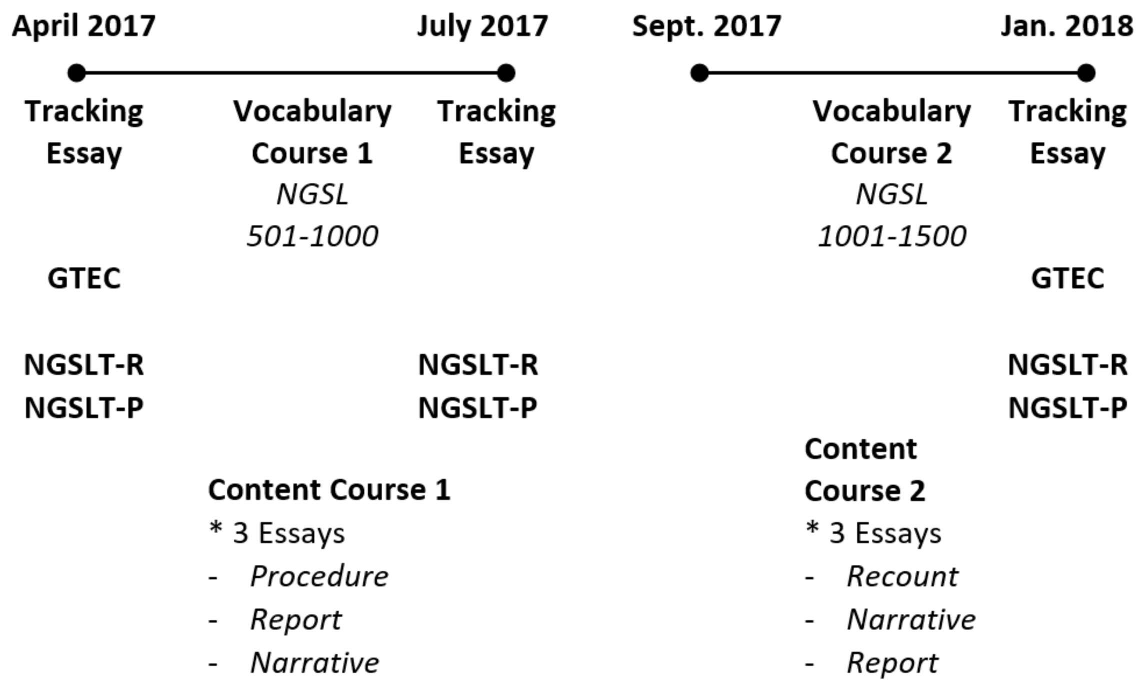

Figure 10. Time schedule showing when the various data collection instruments were administered. Underlined bold indicates tests or essays that were required of all students. Under the name of the courses (italics) are the vocabulary levels covered and the required essays. All essays were required. All vocabulary work and quizzes were tracked. GTEC = Global Test of English Communication, NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive, NGSL = New General Service List.

Quantitative supporting data came in the form of the participants' GTEC, and NGSLT scores, as well as their Quizlet and Moodle self-study metrics, which showed connections between language proficiency, estimated vocabulary size, writing samples, and time spent studying. Furthermore, qualitative data, in the form of interviews, provided insight regarding the students' perceptions of the treatment and helped me triangulate the findings with regard to the salience of the reminders, overall change in English proficiency, and the usefulness of the target vocabulary in their writing.

The participants were placed into one of eight intact vocabulary classes based on their placement test scores and one of 10 content course sections. These classes had previously been assigned to one of the four conditions: three treatment conditions and the control condition. Briefly, the control group did not receive reminders (NR), while each of the treatment groups were given different degrees of reminders: a reminder (R) to use vocabulary, a reminder plus instructions to notice (R+) vocabulary, or a reminder and instructions to use (R++) vocabulary. (See Treatment and Comparison Groups below for a description of the instructions provided to each group).

To ensure that the data collected was representative of the student population, each section of content course 1 was assigned to one of the four conditions before the students were assigned to a particular class. As previously discussed (see Table 13 and Vocabulary Course above), there were 10 sections of content course 1. Two classes were assigned to each of the reminder conditions (R, R+, and R++), and four to the no reminder (NR) control group (see Research Design below). The NR group included two classes from the upper-level block and two from the lower-level block based on placement scores. Each of the treatment conditions was assigned one block at each level. This assignment was done to ensure that each condition was representative of the population of students enrolled. I met with the course instructors before every lesson to confirm with them whether the class should receive the reminder and the degree of reminder or no reminder.

Table 13. *Study Conditions of the Participants (N = 178) by Content Course Section*

	Block A			Block B		
Level	Class	Condition	*n*	Class	Condition	*n*
1 (highest)	1	NR	13	6 (highest)	R++	13
2	2	R	17	7	R+	17
3	3	NR	20	8	NR	20
4	4	R+	18	9	R	20
5 (lowest)	5	R++	21	10 (lowest)	NR	19

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group. Blocks of classes meet at the same time.

This design possesses four strengths. First, because the study was conducted across two consecutive semesters, it was easier to observe changes in vocabulary use and writing quality. Second, varying the strength of the treatment revealed the optimal degree of reminder. Third, the treatment was only administered in the first semester. By continuing the data collection into the second semester, the data from the third Tracking Essay and the various test measures could be analyzed to determine if any of the treatment conditions resulted in continued use of the vocabulary studied in vocabulary course 1 in the content course essays, which would be an indication that long-term retention or learning of the target vocabulary occurred (see Data Analysis below). Finally, the multiple pre- and posttests, combined with the interviews, enabled triangulation of data.

Treatment and Comparison Groups

All the students in the content courses submitted drafts of essays for grading and then revised them before submitting a final draft. The same tasks were administered to classes at different levels to minimize the effect of English proficiency on the results. Table 13 shows the study conditions assigned to each of the content course 1 sections and the number of participants in each class. The participants were divided into two blocks of five classes each; classes 1 through 5, and classes 6 through 10. Students with the highest entry scores were placed in classes 1 and 6, and those with lower entry scores in classes 5 and 10. This placement was similar to the manner used to divide them into the four levels for vocabulary course 1. This resulted in an even, though not identical, distribution of participants to each of the four conditions, with the middle class assigned to the NR (comparison) condition in both blocks.

Comparison condition, no reminder (NR). The students in the NR group (n = 72), Classes 1, 3, 8, and 10, did not receive reminders to use the vocabulary they were studying in their writing assignments; therefore, target vocabulary reflected in their writing was a result of autonomous learning and the subsequent transfer of such vocabulary learning.

Remind condition (R). To investigate the effect of reminders on target vocabulary use, the students in the R group (*n* = 37), classes 2 and 9 in the content course, received a reminder to use recently studied vocabulary (see Table 13). To minimize the effect of instruction or excessive elaboration, the reminder was simply given as, *Please use at least 5 of the 50 words you are studying this week in your homework/writing assignment*. This resulted in the R group receiving the reminder twice a week. To ensure that even low-proficiency students understood the instructions, the reminder was printed on A3-size paper (slightly larger than two U.S. letter-size pages) in English and Japanese and shown to the students when they were told of their homework, and during revisions done in class (see Appendix W).

Remind and notice condition (R+). The R+ group (*n = 35*), classes 4 and 7, followed the same procedure as the R condition, the printed and read reminder, plus the following instructions to notice their vocabulary use. The students in the R+ group were given a list of the vocabulary they were studying that week (see Appendix X) and looked through their drafts to identify words from the list that they had used. Students were instructed to circle or highlight any target words found in their writing.

Remind and use condition (R++). The condition for the R++ group (*n = 34*), classes 5 and 6, followed those of the R group, the printed and read reminder, plus the following instructions to use the vocabulary studied. The students in the R++ group were given a list of the vocabulary that they had studied up until that point, circled words related to the writing topic, and then wrote those words on the paper they used for their final draft (see Appendix Y for the vocabulary list and worksheet received by the R++ participants). The students were then encouraged by the teachers to add those words to their essay when they revised it for submission.

Summary of Research Design

Figure 11 shows the research design of the study and the four conditions. Reminders were only provided in the first semester in Content course 1 for the three reminder conditions: Reminder (R), Reminder plus instructions to notice (R+), and Reminder plus instructions to use (R++). The instruments used to assess the participants' vocabulary levels included a Tracking Essay, administered at three points in the study (entry, end of semester 1, end of semester 2), along with various tests, are discussed in the next section, Instruments.

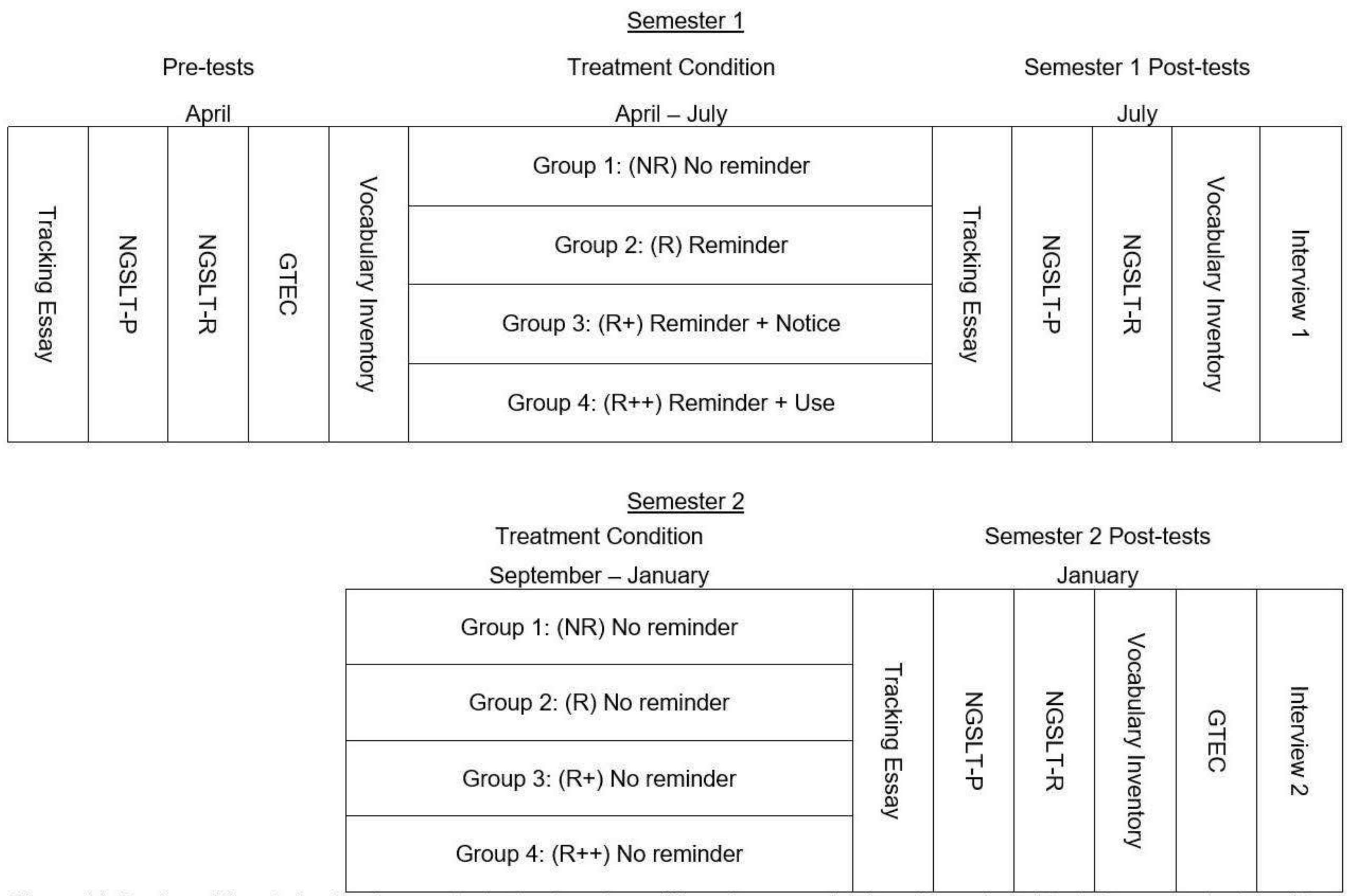

Figure 11. Design of the study showing pre-tests, treatment conditions for semester 1 and 2, and post-tests for semester 1 and 2. NGSLT-P = Productive New General Service List Test, NGSLT-R = Receptive New General Service List Test, GTEC = Global Test for English Communication, TOEIC = Test of English for International Communication.

Data Analysis

In this next section, I detail the data analysis, following the order in which measures were given (see Figure 10 for a timeline of administration and Figure 11 for an overview of the design of the study). The study began in April, 2017, the start of the academic year in Japan, and ran until January the following year.

Having outlined the research design, the instruments that were used, and when the data were collected, in the next section I describe the data analysis, beginning with Text Lex Compare and the Chi-square procedure. Following this I describe the analysis of essay quality and the analysis of used versus unused vocabulary in the participants' writing. I conclude by detailing the analysis of the qualitative data.

Text Lex Compare. As the primary focus of this study was target vocabulary use in writing, the main source of datum was the tracking essays. The essays were transcribed by me and saved as text documents. For the essays collected, the analysis compared the target vocabulary with groups of essays (e.g., all #1 tracking essays (TE1) against all #2 tracking essays (TE2)), and also the essays for individual students (e.g., Student A's TE1 against Student A's TE2) to show the target words that were used in writing. An analysis was also run to investigate if any of the target words were already used by the students at the start of the study. It was hypothesized that vocabulary use reminders would be more salient in stronger treatments which would lead to more time spent studying (hypothesis 1), improved vocabulary learning and retention (hypothesis 2), improved GTEC and writing scores (hypothesis three), more correct use of the target vocabulary (hypothesis four), and that the participants who were reminded would report greater perceived usefulness of the vocabulary study as it relates to their writing (hypothesis 5).

Target vocabulary used in the tracking essays was checked using the Text Lex Compare online resource (Cobb, n.d.-a). Text Lex Compare analyzes two or more texts and shows which words are repeated in those texts. The output of the program includes the number of tokens and word families in the texts as well as words that are unique to each of the texts. By submitting a participant's Tracking Essay and a relevant vocabulary list, use of target vocabulary could be identified.

The analyses of the tracking essays using Text Lex Compare was used to explore hypothesis two, *increased vocabulary study time will improve vocabulary learning and retention for the reminder groups*, and hypothesis three, *the participants receiving stronger treatments will be able to use more target vocabulary than those in weaker treatments*. Text Lex Compare (Cobb, n.d.-a) was used to analyze three aspects of target vocabulary use in the tracking essays. First, all TE1 data were compared to the first 2,000 words of the NGSL to examine target vocabulary use at the outset of the study. Second, all TE2 data were compared to the first 1,000 words of the NGSL and TE1 to determine the effect that reminders and reminder conditions had on target vocabulary use after one semester. Finally, all TE3 data were compared to TE1 and TE2, and the first 2,000 words of the NGSL to examine the change in target vocabulary use after two semesters and evaluate if any of the reminder conditions (R, R+, R++) had a lasting effect when compared with the no reminder (NR) condition or the other two treatment conditions.

After this, a Chi-square test of independence was run to evaluate if differences between the NR, R, R+, and R++ groups were significant. Effect sizes were also calculated. A key assumption of the Chi-square test is that each participant only contributes data to one cell. This means that a participant's score for a particular Tracking Essay was used for analysis only once in the Chi-square procedure. Other assumptions of the Chi-square procedure were checked and met. These assumptions were that the data in the cells were frequencies, the categories of the variables were mutually exclusive, and there was only one categorical variable (treatment type). The Chi-square test was used to explore hypothesis four, *improved vocabulary learning and retention for the reminder groups will enable the participants to use the target vocabulary more accurately in the tracking essays*. Correct usage was where there were no major spelling mistakes and comprehension was not impeded. Partially correct usage was where the usage was contextually inappropriate with no impediment to comprehension. Incorrect usage was where the participant displayed a lack of understanding of the word and the written intent was not successfully conveyed.

The effect sizes were also calculated and the confidence intervals showed whether effect sizes of practical importance were still concordant with the data. The effect size was calculated using the following formula (Field, 2013):

$$d = (M_1 - M_2) / SD$$

where $M_1 - M_2$ are the means for the 1st and 2nd samples, and SD is the standard deviation for the samples.

Correlation analyses were conducted between the GTEC, cumulative NGSLT-Receptive and NGSLT-Productive scores, writing scores, self-study time (as measured by Quizlet, 2017) and self-test time (as measured by Moodle, 2016) to evaluate their relationships. It was anticipated that cumulative NGSLT-Receptive & NGSLT-Productive scores would correlate positively with the results from the Text Lex Compare analysis because it was hypothesized that stronger treatments would result in more self-study time, which would translate to greater scores on the vocabulary tests and tracking essays. The correlation analyses was used to explore hypothesis one, *the reminders will result in more time being spent on studying the target vocabulary using Quizlet and Moodle*, hypothesis two, *vocabulary study time is expected to correlate positively with the vocabulary test* *scores*, and hypothesis three, *improved vocabulary learning and retention for the reminder groups will lead to better GTEC and tracking essay scores*.

Vocabulary inventory survey. The participants' vocabulary inventory surveys consisted of words from the second 500 words of the New General Service List (NGSL) (Browne et al., 2013b). The surveys were examined to determine the words that were known and unknown by the participants at the start of the first semester in April, at the end of the first semester in July, and at the end of the second semester in January The total number of known words reported for each participant was recorded and used for comparison. This analysis revealed if short-term and/or long-term learning was occurring and provided insights into the changes in the participants' target vocabulary use. This analysis contributed to interpreting the data in exploring hypothesis three, *the participants receiving stronger treatments will be able to use more target vocabulary than those in weaker treatments,* and *vocabulary study time is expected to correlate positively with the vocabulary test scores,* and hypothesis four, *participants in stronger treatments will demonstrate more correct usage of the target vocabulary*.

Writing quality. The participants' tracking essays were evaluated using three tools: a rubric designed by the Writing Research Group of the JACET Kansai Chapter (hereafter JACET writing rubric) (Kuru, Masaki, & Kinshi, 2009), an analysis of lexical complexity (Ong & Zhang, 2010), and a count of standard length words (Carver, 1982). These analyses were used to explore hypothesis four, *the higher incidence of correct vocabulary use will lead to better overall writing scores (JACET writing rubric, lexical complexity, standard length words)*.

The first evaluation of the participants' writing was conducted using the JACET writing rubric (Kuru et al., 2009) to investigate if their writing quality progressed throughout the two semesters. This rubric was selected over a more well-known rubric, the ESL Composition Profile (Hughey, Wormuth, Hartfiel, & Jacobs, 1983), as the latter was not designed for an EFL environment with a specific first language (i.e., Japanese). One rater evaluated all essays, but a 10% sample was evaluated by a second rater to establish interrater reliability. These data were analyzed using FACETS (Linacre, 2014). The essays were analyzed using the formula WT^2/W (word types squared divided by the total number of words) (Ong & Zhang, 2010). This formula for lexical complexity was chosen as it takes into account the varying lengths of the texts collected, avoids the problems with the formula for lexical variation (Laufer, 2003a) which is both unstable for short texts (i.e., less than 200 words) and affected by differences in text length.

The third evaluation was a count of standard length words of each participant's tracking essays. Carver (1982) defined a standard length word as containing 6 letter-spaces which includes letters, spaces, and punctuation. For example, the sentence *This study was conducted in Japan* consists of 33 letter-spaces which equates to 5.5 standard length words. Counting the number of standard length words in a text is more accurate than counting individual words because one word can be significantly longer or shorter than another word. Using standard length words was, therefore, one way of measuring writing fluency and made the writing produced by one student more easily comparable to that produced by others.

Qualitative data. Qualitative data were collected from the interviews that were conducted at the end of each semester. These data were subjected to interpretive analysis (Hatch, 2002, p. 181). The steps for interpretative analysis are shown in Figure 12 (Hatch, 2002, p. 181). These qualitative data were intended to support the quantitative data and was used to explore hypothesis one, *the reminders will be more salient in the stronger treatments*, and hypothesis 5, *learners who were reminded will report greater perceived usefulness of the vocabulary study as it relates to their writing*. These hypotheses focused on the participants' perceptions of the treatment including the salience of the reminders, usefulness of vocabulary in writing, improvement in English level, effects of the reminders on their writing, and vocabulary knowledge.

1. Read the data for a sense of whole.
2. Review impressions previously recorded in research journals and/or bracketed in protocols, and record these in memos.
3. Read the data, identify impressions, and record impressions in memos.
4. Study memos for salient interpretations.
5. Reread the data, coding places where interpretations are supported or challenged.
6. Write a draft summary.
7. Review interpretations with participants.
8. Write a revised summary and identify excerpts that support interpretations.

Figure 12. Steps in Interpretive Analysis. Adapted from Doing Qualitative Research in Education Settings (p. 181), by J. Amos Hatch, 2002, New York: State University of New York Press. Copyright 2002 by State University of New York Press. Used with permission.

Rasch Analysis

A Rasch analysis was conducted using WINSTEPS, version 3.80.0 (Linacre, 2013) to conduct preliminary analyses of the receptive (NGSLT-Receptive) and productive (NGSLT-Productive) vocabulary tests. validity evidence was gathered using the dichotomous Rasch model. In addition, the Rasch analysis was used to convert the raw scores into interval Rasch logits.

An additional Rasch software package, FACETS, version 3.71.4 (Linacre, 2014), was used to evaluate the raters, the written characteristics of the tracking essays, and to produce fair averages of the participants' writing performance. A fair average is the expected observed score if the candidate had been rated by an unbiased average severity judge. FACETS accomplishes this by calibrating the examinees, raters, tasks, and rating scales onto the same interval scale expressed in logits (logarithm of odds units). This interval scale creates a single frame of reference for interpretation of the results (Eckes, 2005). The many-faceted Rasch model makes it possible assess examinee performance based on a number of facets. In this study, the five facets of the tracking essays were content and idea development, organization, grammar, vocabulary, and mechanics.

Rasch Fit Criteria

The output from WINSTEPS and FACETS was used to evaluate how well the items on the NGSLT-Receptive and NGSLT-Productive, as well as the participants, fit the Rasch model. Rasch analysis compares a real world data set to a theoretical model; significant discrepancies between the two warrant closer investigation. For example, participants who misfit the model might not have responded to items seriously or truthfully. In the context of this study, fit problems might be caused by carelessness or guessing. Items that misfit the model indicate potential item bias or redundancy. Raters often misfit the Rasch model because of bias or a misunderstanding or misuse of the rating criteria. Meeting the various fit criteria is one key aspect of Rasch analysis.

Person and item fit were determined through an examination of the infit and outfit statistics. The infit statistic is the standardized residual weighted by individual variance (W_{ni}) and is calculated using the formula

$$\text{Infit} = \Sigma Z_{ni}^2 W_{ni} \,/\, \Sigma W_{ni}$$

The outfit statistic is the unweighted average of the standardized residual (Z_{ni}) variance across both persons and items and is calculated using the formula

$$\text{outfit} = \Sigma Z_{ni}^2 \,/\, N$$

Both unstandardized and standardized fit statistics were examined. The unstandardized mean fit square statistics indicate the magnitude of misfit between the person or item and the model (Bond & Fox, 2007). A value of 1.00 shows that the person or item fits the model perfectly (Linacre, 2002). Values < 1.00 indicate overfit, which signifies that persons or items are behaving too predictably. Overfit results in a reliability and separation estimates that can be misleadingly high with lower error variances. Values > 1.00 indicate underfit and affects the precision of the measurement of the latent variable (Linacre, 2002). Underfit is, therefore, a potentially greater risk than overfit. Standardized fit statistics represent the likelihood of misfit (Bond & Fox, 2007). The standardized values are represented as z-scores with an expected value of 0.00. ZSTD values outside the range of -2.00 to 2.00 correspond to a Type I error rate of .05.

The infit statistics are sensitive to inliers. Infit statistics for items are focused on unexpected responses by persons whose ability estimates are close to a particular item's difficulty estimate. Conversely, infit statistics for persons are focused on unexpected responses to items whose difficulty estimate is close to that particular person's ability estimate. In contrast to the infit statistics, outfit statistics are sensitive to outliers, which means that they are sensitive to unexpected responses by persons on items that are relatively far above or below their ability or difficulty estimate (Linacre, 2002). Wright, Linacre, Gustafson, and Martin-Lof (1994) have suggested that fit values between .70 to 1.30 are considered productive for measurement for multiple-choice tests.

Rasch Reliability and Separation

Rasch output from WINSTEPS and FACETS reports reliability and separation statistics for persons and items. A high reliability means that there is a high probability that persons or items with high estimated measures actually have higher measures than persons or items with lower estimated measures. Essentially, high person reliability requires a wide range of ability among the participants and/or an instrument with many items. Similarly, high item reliability requires a test with a wide range of difficulty and/or a large sample of persons (Linacre, n.d.-a). Both person and item reliability statistics give an indication of the reproducibility of the data. Person or item reliability is calculated using the formula

(OV – EV) / EV

where OV is the observed variance of person ability or item difficulty measures and EV is the mean of squared standard errors of person ability or item difficulty measures (Linacre, n.d.-a).

The separation statistics are used to estimate the spread of persons and items on the measured variable. Low person separation indicates that the instrument may not be sensitive enough to distinguish between individuals in the sample population. Similarly, low item separation implies that the sample population is not large enough to confirm the item difficulty hierarchy (Linacre, n.d.-a). Person or item separation is calculated using the formula

Separation = True SD / RMSE

where True SD is the standard deviation of reported measures corrected for measurement error inflation and RMSE is the average measurement error of reported measures (Fisher, 1992).

For this study, reliability and separation criteria proposed by Fisher (2007) were used. Reliability estimates below .67 are considered poor, .67 to .80 are fair, .81 to 90 are good, .91 to .94 are very good, and values greater than .94 are excellent. Separation estimates below 2.00 are poor, 2.00 to 3.00 are fair, 3.00 to 4.00 are good, 4.00 to 5.00 are very good, and values greater than 5.00 are excellent.

Rasch Principal Components Analysis of Item Residuals

One key assumption of the Rasch model is that the data are unidimensional (Bond & Fox, 2007). Indeed, the ideal Rasch model would be one where all the information in the data is explained by the latent measures. Any unexplained part of the data, denoted by the residuals, should be random. However, if analysis of the residuals reveals a systematic pattern, this suggests the presence of a secondary dimension which would distort the measurement. A Rasch principal components analysis (PCA) of the item residuals looks for systematic patterns in the unexpected parts of the data to verify that the data are fundamentally unidimensional. For this study, the dimensionality of the instruments was analyzed using the following criteria (Linacre, 2007):

1. Variance explained by items: Greater than four times the first contrast is good.
2. Variance explained by measures: Greater than 50% is good.
3. Unexplained variance explained by first contrast: Less than 3.0 is good; Less than 1.5 is excellent.
4. Unexplained variance explained by first contrast: Less than 5% is excellent.

The primary focus is on criteria 3 and 4 as a second construct, if present, will be revealed in the contrasts.

Summary

The purposes of this chapter were to introduce the context of the study, including the setting, provide details about the courses from which participants were drawn, provide information about study participants, outline the methodological framework, to present the instruments, describe the methods used to analyze the data, and to explain the rationale behind the approaches used to analyze the results. The purposes of Chapter 4 are to describe two pilot studies that informed the data-collection procedures and research design of this study.

CHAPTER 4

PRELIMINARY ANALYSES

The purpose of this chapter is to report on the preliminary analyses of the receptive and productive vocabulary tests, the NGSLT-Receptive and NGSLT-Productive, and the tracking essays.

Receptive Vocabulary Test (NGSLT-Receptive)

The NGSLT-Receptive was examined using the Rasch model (Rasch, 1960) with WINSTEPS, version 3.80.0 (Linacre, 2013). The raw scores from the NGSLT-Receptive were stacked and analyzed to incorporate the participants' scores from Time 1, Time 2, and Time 3. Stacking the data produced an "average" underlying rating scale of the data derived from the three rounds of testing (Wright, 2003). The item difficulty along with the spread of item difficulty estimates was checked as well as the item fit, dimensionality, item and person reliability, and separation indices. Table 14 shows the item difficulty estimates. The items ranged from -3.28 to 3.39 logits. The item that was rated to have the least difficulty was item 28, *global*. The item that was rated most difficult was item 47, *comprehensive*. Overall, the NGSLT-Receptive had a Rasch item reliability estimate of 0.98, which is considered excellent according to Fisher (2007). The Rasch item separation was 6.88. The Rasch person reliability estimate was 0.74, which is considered fair. The person separation statistic was 1.71, which implies that the instrument may not be sensitive enough to distinguish between high and low performers (Linacre, n.d.-a).

All the items met the infit mean square criteria of 0.70 to 1.30 (Wright et al., 1994), as the infit mean square statistics ranged from 0.84 to 1.26. Even though the standardized mean square statistics indicated that items 47, 4, 3, 42, and 31 had values above 2.0, and items 36 and 44 had an infit less than -2.0, these values can be ignored if the mean square statistics meet the fit criteria (Linacre, n.d.-b).

Items 27, 14, 13, 2, 47, 23, 8, and 5 did not meet the outfit mean square criteria and had values ranging from 1.40 to 1.86 (see Table 14). These values indicate that the items were underfitting the model suggesting that these items were the result of a few random responses by low performers (Linacre, n.d.-b). Items 20, 37, 33, 12, 19, 7, 39, 36, and 10 had values ranging from 0.57 to 0.69 and were overfitting the model suggesting a lack of item independence.

The item-person Wright map displayed in Figure 13 shows that the item difficulty was relatively low compared to estimated person ability. The most difficult item, Q47 *comprehensive*, was below the estimated ability of roughly 20% of the participants. Furthermore, 11 items were redundant as they were below the lowest estimated ability of the participants.

Table 14. *Rasch Item Statistics for the 50-Item NGSLT-Receptive*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
47	3.39	0.10	1.26	5.30	1.62	6.70	0.18
18	2.87	0.10	1.05	1.40	1.14	2.20	0.38
43	1.87	0.10	0.94	-1.40	0.94	-0.90	0.47
42	1.74	0.10	1.09	2.20	1.11	1.70	0.34
44	1.68	0.10	0.87	-3.20	0.84	-2.50	0.52
45	1.67	0.10	0.99	-0.30	0.97	-0.50	0.43
34	1.51	0.10	1.06	1.30	1.04	0.60	0.36
31	1.40	0.10	1.10	2.10	1.17	2.00	0.32
3	1.40	0.11	1.12	2.40	1.12	1.40	0.31
11	1.34	0.11	0.91	-1.90	0.84	-2.00	0.48
4	1.31	0.11	1.20	3.80	1.26	2.90	0.23
46	1.29	0.11	0.91	-1.90	0.81	-2.30	0.48
24	1.29	0.11	0.91	-1.70	0.87	-1.50	0.47
50	1.25	0.11	0.90	-1.90	0.82	-2.10	0.48
25	1.19	0.11	0.94	-1.10	0.82	-2.00	0.45
17	1.15	0.11	0.90	-1.90	0.83	-1.90	0.47
48	0.84	0.12	0.91	-1.50	0.77	-2.20	0.46
36	0.81	0.12	0.84	-2.60	0.67	-3.20	0.51
49	0.71	0.12	1.07	1.10	1.09	0.80	0.30
13	0.43	0.13	1.08	1.00	1.49	3.00	0.23
29	0.37	0.13	1.02	0.30	0.99	0.00	0.32
32	0.20	0.14	0.96	-0.40	0.90	-0.60	0.34
27	0.12	0.14	1.13	1.30	1.40	2.10	0.17
7	0.08	0.14	0.87	-1.40	0.65	-2.20	0.43
16	0.04	0.14	1.05	0.50	0.95	-0.20	0.27
15	0.00	0.15	0.95	-0.50	0.83	-0.90	0.35
1	-0.04	0.15	1.05	0.50	1.16	0.90	0.25
41	-0.11	0.15	0.92	-0.70	0.70	-1.60	0.37
2	-0.16	0.15	1.11	1.10	1.61	2.60	0.16
22	-0.16	0.15	0.96	-0.30	0.92	-0.30	0.31
9	-0.36	0.16	0.99	0.00	0.90	-0.40	0.27
14	-0.38	0.17	1.06	0.50	1.41	1.70	0.18
35	-0.72	0.19	0.99	0.00	1.05	0.30	0.23
10	-0.87	0.20	0.98	0.00	0.69	-1.10	0.27
26	-0.87	0.20	1.00	0.10	0.92	-0.20	0.23
33	-1.00	0.21	0.91	-0.50	0.64	-1.20	0.31
8	-1.04	0.21	1.07	0.40	1.72	2.00	0.11
30	-1.09	0.22	1.01	0.10	1.01	0.10	0.19
19	-1.24	0.23	0.97	-0.10	0.65	-1.10	0.25
23	-1.29	0.23	0.99	0.00	1.72	1.90	0.17
6	-1.35	0.24	0.99	0.00	1.29	0.90	0.17
37	-1.35	0.24	0.94	-0.20	0.63	-1.10	0.25
39	-1.35	0.24	0.99	0.00	0.67	-1.00	0.22
20	-1.47	0.25	0.94	-0.20	0.57	-1.30	0.26
21	-2.03	0.33	0.94	-0.10	0.70	-0.60	0.19
38	-2.15	0.34	0.97	0.00	1.09	0.40	0.14
40	-2.15	0.34	0.98	0.10	1.02	0.20	0.17
5	-2.76	0.45	1.02	0.20	1.86	1.50	0.06
12	-2.76	0.45	0.96	0.10	0.65	-0.60	0.14
28	-3.28	0.58	0.95	0.10	1.07	0.30	0.11

```
----------------------------------------------------------------------
MEASURE    Person - MAP - Item
               <more>|<rare>
    5             #  +
               .###  |
                     |
                    T|
                     |
                     |
    4        .#####  +
                  .  |
                     |
          .######## S|
                     |  Q47
        .##########  |
    3                +T
         .#########  |  Q18
         ##########  |
           ######## M|
            .######  |
                  .  |
    2        .#####  +
        .##########  |  Q43
              .####  |  Q42    Q44    Q45
               .###  |S Q34
               .### S|  Q11    Q24    Q3     Q31    Q4     Q46
                 ##  |  Q17    Q25    Q50
    1          .###  +
                 .#  |  Q36    Q48
                .##  |  Q49
                  .  |  Q13
                .## T|  Q29
                  .  |  Q27    Q32
    0             .  +M Q1     Q15    Q16    Q7
                  .  |  Q2     Q22    Q41
                  .  |  Q14    Q9
                     |
                     |  Q35
                     |  Q10    Q26
   -1                +  Q33    Q8
                  .  |  Q19    Q30
                     |  Q23    Q37    Q39    Q6
                     |S Q20
                     |
                     |
   -2                +  Q21
                     |  Q38    Q40
                     |
                     |
                     |
                     |  Q12    Q5
   -3                +T
                     |
                     |  Q28
                     |
               <less>|<frequent>
----------------------------------------------------------------------
```

Figure 13. Item-person Wright map for the New General Service List Test-Receptive. Each # equals 5 persons. Each . equals 1 to 4 persons. M = Mean; S = One standard deviation from the mean; T = Two standard deviations from the mean.

The dimensionality of the NGSLT-Receptive test was investigated using a Rasch principal component analysis of item residuals. Dimensionality was checked using the four criteria outlined in Linacre (2007). As seen in Table 15, the variance explained by the items was 15.70%. This percentage was greater than four times the variance accounted for by the first contrast (2.60%); thus, the first criterion was met. The variance explained by the measures was 28.10% (eigenvalue = 19.6), which did not meet the second > 50% criterion. The third criterion was also met because the first residual contrast had an eigenvalue of 1.8 which is considered good. This eigenvalue represented 2.60% of the variance which is considered excellent and met the fourth criteria.

Table 15. *Item Dimensionality (PCA of item residuals) for the NGSLT-Receptive*

	Empirical		Modeled
	Eigenvalue	Percentage	Percentage
Total raw variance in observations	69.6	100.00%	100.00%
Raw variance explained by measures	19.6	28.10%	28.20%
Raw variance explained by persons	8.6	12.40%	12.40%
Raw Variance explained by items	10.9	15.70%	15.80%
Raw unexplained variance (total)	50.0	71.90%	71.80%
Unexplained variance in 1st contrast	1.8	2.60%	3.60%

Note. NGSLT-Receptive = New General Service List Test-Receptive.

The Rasch item reliability estimate at Time 1 was 0.93, which was very good (Fisher, 2007). The Rasch item separation index was 3.67. The Rasch person reliability estimate was .78 which is considered fair and the person separation statistic was 1.87.

At Time 2, the Rasch item reliability estimate was .94, and the item separation index was 3.88. The Rasch person reliability estimate was .74 and the person separation statistic was 1.70.

At Time 3, the Rasch item reliability estimate was .88, and the item separation index was 2.73. The Rasch person reliability estimate was .67 and the person separation statistic was 1.44.

Productive Vocabulary Test (NGSLT-Productive)

The NGSLT-Productive was examined using the dichotomous Rasch model (Rasch, 1960) with WINSTEPS, version 3.80.0 (Linacre, 2013). The raw scores from the NGSLT-Receptive were stacked and analyzed to incorporate the participants' scores from Time 1, Time 2, and Time 3. The item difficulty along with the spread of item difficulty estimates was checked as well as the item fit, dimensionality, item and person reliability, and separation indices. Table 16 shows the item difficulty estimates. The items ranged from -2.93 to 2.91 logits. The item that was rated to have the least difficulty was item 23, *sad*. The item that was rated most difficult was item 20, *impose*. Overall, the NGSLT-Productive had a Rasch item reliability estimate of 0.99, which is considered excellent according to Fisher (2007). The Rasch item separation was 13.01. The Rasch person reliability estimate was 0.81, which is considered good. The person separation statistic was 2.08.

Table 16. *Rasch Item Statistics for the 30-Item NGSLT-Productive at Time 1*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
20	2.91	0.13	1.03	0.40	1.07	0.50	0.40
22	2.65	0.13	0.92	-1.10	0.77	-1.60	0.50
30	2.54	0.12	0.85	-2.10	0.74	-2.00	0.54
14	2.45	0.12	1.28	3.80	1.52	3.40	0.25
29	2.21	0.12	0.95	-0.70	0.84	-1.40	0.50
27	2.08	0.11	0.91	-1.60	0.82	-1.80	0.53
25	1.66	0.11	0.98	-0.40	1.07	0.80	0.48
15	1.56	0.10	0.91	-1.90	0.87	-1.70	0.53
28	1.24	0.10	0.97	-0.70	0.93	-1.00	0.50
26	1.12	0.10	0.89	-2.60	0.84	-2.50	0.55
18	0.70	0.10	0.81	-5.30	0.74	-4.30	0.60
10	0.55	0.10	1.12	3.00	1.15	2.10	0.38
6	0.43	0.10	1.04	1.10	1.04	0.60	0.43
13	0.40	0.10	0.98	-0.50	0.94	-0.80	0.47
4	-0.16	0.10	1.08	1.90	1.18	1.90	0.36
21	-0.39	0.11	1.05	1.00	1.19	1.70	0.37
11	-0.59	0.11	1.20	3.50	1.57	4.10	0.24
17	-0.64	0.11	1.07	1.20	1.21	1.60	0.34
8	-0.68	0.11	0.90	-1.90	0.82	-1.50	0.47
16	-0.83	0.11	1.09	1.50	1.18	1.30	0.31
19	-1.30	0.13	1.00	0.00	0.98	-0.10	0.34
5	-1.32	0.13	1.08	1.10	1.01	0.10	0.29
2	-1.45	0.13	0.91	-1.00	0.74	-1.50	0.40
9	-1.49	0.13	0.97	-0.30	0.81	-1.00	0.35
24	-1.51	0.14	1.03	0.40	1.22	1.20	0.29
12	-1.92	0.15	0.95	-0.40	0.77	-1.00	0.33
7	-2.35	0.18	0.98	-0.10	1.17	0.70	0.24
3	-2.41	0.18	0.97	-0.20	0.74	-0.90	0.28
1	-2.51	0.19	0.99	0.00	0.77	-0.80	0.26
23	-2.93	0.22	1.04	0.30	0.70	-0.90	0.20

Note. NGSLT-Productive = New General Service List Test-Productive.

All the items met the infit mean square and outfit mean square criteria of 0.70 to 1.30 (Wright et al., 1994), as the infit mean square statistics ranged from 0.81 to 1.28.

The outfit mean square statistics ranged from 0.70 to 1.57 with items 14 and 11 having values of 1.52 and 1.57, respectively. Even though the standardized mean square statistics indicated that item 10 had a value above 2.0, this value can be ignored if the mean square statistics meet the fit criteria (Linacre, n.d.-b). The point-measure correlations were all positive and ranged from 0.20 to 0.60.

The item-person Wright map displayed in Figure 14 shows that the item difficulty was spread evenly across the participants' ability level. The most difficult item, *impose*, was at or below the estimated ability of approximately 10% of the participants and one participant's estimated ability was below the easiest item, *sad*.

The dimensionality of the NGSLT-Productive was investigated using a Rasch principal component analysis of item residuals. The dimensionality was checked using the four criteria outlined in Linacre (2007). As seen in Table 17, the variance explained by the items was 27.30%, which was greater than four times the variance accounted for by the first contrast (3.70%). Thus, the first criterion was met. The variance explained by the measures was 43.40% (eigenvalue = 23.0). This figure did not meet the second criterion of > 50%. The third criterion was met because the first residual contrast had an eigenvalue of 1.9 which is considered good. This eigenvalue represented 3.70% of the variance, which is considered excellent.

```
---------------------------------------------------------------------
MEASURE    Person - MAP - Item
               <more>|<rare>
    5             .  +
                     |
                     |
                  #  |
                     |
                     |
    4                +
                     |
                .##  |
                     |T
                     |
               .### T|
    3                +
                .##  |  impose
                     |  justify
                ###  |  accurate    extent
                     |
            #######  |  bias        throat
    2              S+
             ######  |
         .#########  |S motivation
                     |  guarantee
        .##########  |
           ########  |  diversity   gradually
    1                +
          .########  |
        .########## M|  sheet
                     |  benefit     consider
      .############  |  extra
          .########  |
    0      .#######  +M
                     |  include
          .########  |  curve
             .##### S|
               ####  |  examine     global      solution
              .####  |  factory
   -1                +
                ###  |
               .###  |  steal       well
                     |  pressure    rise        shadow
                  # T|S
                  .  |
   -2                +  deep
                  .  |
                     |  fly         health
                  .  |  different
                     |
                     |
   -3                +  sad
                  .  |
                     |
                     |T
               <less>|<frequent>
---------------------------------------------------------------------
```

Figure 14. Item-person Wright map for the NGSLT-Productive. Each # equals 4 persons. Each . equals 1 to 3 persons. M = Mean; S = One standard deviation from the mean; T = Two standard deviations from the mean.

Table 17. Item Dimensionality (PCA of Item Residuals) for the NGSLT-Productive

	Empirical		Modeled
	Eigenvalue	Percentage	Percentage
Total raw variance in observations	53.0	100.00%	100.00%
Raw variance explained by measures	23.0	43.40%	43.30%
Raw variance explained by persons	8.5	16.10%	16.10%
Raw Variance explained by items	14.5	27.30%	27.30%
Raw unexplained variance (total)	30.0	56.60%	56.70%
Unexplained variance in 1st contrast	1.9	3.70%	6.50%

Note. PCA = principle component analysis, NGSLT-Productive = New General Service List Test-Productive.

At Time 1, the Rasch item reliability estimate was 0.98, and this is excellent according to Fisher (2007). The Rasch item separation index was 7.35. The Rasch person reliability estimate was 0.73 which is considered fair. The person separation statistic was 1.63 which implies that the instrument may not be sensitive enough to distinguish between high and low performers (Linacre, n.d.-a).

At Time 2, the Rasch item reliability estimate was .98, and the item separation index was 6.97. The Rasch person reliability estimate was .80 and the person separation statistic was 2.00.

At Time 3, the Rasch item reliability estimate was .96, and the item separation index was 4.78. The Rasch person reliability estimate was .77 and the person separation statistic was 1.83.

Tracking Essays

The tracking essays were examined using the many-faceted Rasch model (Rasch, 1960) with FACETS, version 3.71.4 (Linacre, 2014). The raw scores using the JACET writing rubric (Kuru et al., 2009) were stacked and analyzed to incorporate the participants' essay scores from Time 1, Time 2, and Time 3. The fit and reliability of the sessions, rating criteria, and the category statistics are examined and reported.

The results of the FACETS analysis for each session of the tracking essays is shown in Table 18. The measures indicated that the mean Rasch difficulty estimates decreased from 0.74 at Time 1 to -0.04 at Time 2, and then to -0.70 at Time 3. This shows that the tracking essays were getting progressively easier. The infit mean square ranged from 0.92 to 1.13 and outfit mean square from 0.89 to 1.14. Thus, the fit criterion of 0.70 to 1.30 was met.

Table 18. *Session Measurement Report of the Tracking Essays*

Session	Obsvd. Average	Fair Average	Measure	Model SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PtMea
1	2.36	2.58	0.74	0.07	1.13	2.90	1.14	2.90	0.85
2	2.80	2.78	-0.04	0.07	0.94	-1.30	0.94	-1.20	0.85
3	3.19	2.93	-0.70	0.07	0.92	-1.90	0.89	-2.30	0.85

Note. Obsvd. Average = Observed Average, PtMea = Point-measure correlations.

The mean Rasch item difficulty estimates for each rated component of the tracking essays ranged from -0.39 to 0.48 (see Table 19). Vocabulary had the highest difficulty estimate followed by grammar, mechanics, organization, and content. Thus, linguistic components (vocabulary and grammar) were the most difficult criteria on which to get a high score with vocabulary being the most difficult. Content and Idea Development was the easiest criterion. All items met the infit mean square criteria.

Table 19. *Item Measurement Report*

Rating Criteria	Obsvd. Average	Fair Average	Measr	Model SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-Mea
Vocab.	2.67	2.65	0.48	0.09	0.78	-4.2	0.76	-4.1	0.89
Gramr.	2.75	2.75	0.10	0.09	0.86	-2.6	0.85	-2.5	0.88
Mech.	2.76	2.75	0.08	0.09	1.19	3.2	1.19	2.9	0.84
Org.	2.83	2.84	-0.27	0.09	0.95	-0.8	0.92	-1.2	0.88
CID	2.86	2.86	-0.39	0.09	1.20	3.3	1.23	3.5	0.86

Note. Vocab. = Vocabulary, Gramr. = Grammar, Mech. = Mechanics, Org. = Organization, CID = Content & Idea Development, Obsvd. Average = Observed Average, Measr = Measure, Pt-Mea = Point-Measure Correlations.

Rasch rating category statistics for the tracking essays are shown in Table 20. In general, the categories functioned well according to four rating scale diagnostic criteria: (a) category, (b) average measures, (c) category fit, (d) threshold estimates. Examining the category frequency, the shape of the distribution was positively skewed with more ratings at the lower end of the scale. Inspecting the initial count data, all ratings met the minimum number of 10 responses (Linacre, 1999).

The average measures for each rating category, which ranged from -6.68 to 6.16, increased evenly through the rating categories. Therefore, the rating category functioned as expected. In addition, each rating category met the outfit mean square criterion less than 2.00 (Bond & Fox, 2007). Because the distances between adjacent threshold estimates were between 1.4 to 5 logits (Eckes, 2015), they met this recommended criterion. Finally, the standard error was low reflecting a reasonably accurate scale.

Table 20. *Category Statistics for the Tracking Essays*

Rating Category	Descriptor	Count (%)	Average Measure	Outfit MNSQ	Rasch-Andrich Thresholds	SE	Rash-Thurstone Thresholds
1	Very poor	259 (8%)	-6.68	1.0			Low
2	Poor	885 (31%)	-3.70	1.0	-6.69	.09	-6.69
3	Average	1094 (38%)	-0.21	1.0	-2.13	.06	-2.13
4	Good	603 (21%)	3.14	1.0	2.02	.07	2.02
5	Very Good	89 (3%)	6.16	0.9	6.79	.14	6.79

The FACETS map for the JACET writing Rubric categories is shown in Figure 15. The first column is the logit scale. The second column displays the participants' Rasch ability estimates across the logit scale. The third column shows the rater severity estimates with Rater 2 being more lenient than Rater 1. The fourth column shows the times at which the participants wrote the tracking essays and indicates that the essays were getting easier. The fifth column shows the difficulty estimates of the rating components. Vocabulary was the most difficult component to endorse while content and idea development, and organization were the easiest. The sixth column shows the rating scale with a score of 5 being the highest and 0 being the lowest.

```
+-----------------------------------------------------------------------+
|Measr|+examinee   |-Rater   |-Session |-Item                    |Scale|
|-----+------------+---------+---------+-------------------------+-----|
|   8 + .          +         +         +                         + (5) |
|     |            |         |         |                         |     |
|   7 + *.         +         +         +                         + --- |
|     |            |         |         |                         |     |
|   6 + *          +         +         +                         +     |
|     | *.         |         |         |                         |     |
|   5 + *.         +         +         +                         +     |
|     | *.         |         |         |                         |  4  |
|   4 + *.         +         +         +                         +     |
|     | ***        |         |         |                         |     |
|   3 + ***        +         +         +                         +     |
|     | **         |         |         |                         |     |
|   2 + *****.     +         +         +                         + --- |
|     | ********.  |         |         |                         |     |
|   1 + .          +         +         +                         +     |
|     | ********.  | Rater2  | April   | Vocabulary              |     |
*   0 * **.        *         * July    * Grammar   Mechanics     *  3  *
|     | ****.      | Rater1  | January | Content   Organization  |     |
|  -1 + ******.    +         +         + & Idea Development      +     |
|     | ******.    |         |         |                         |     |
|  -2 + *******.   +         +         +                         + --- |
|     |            |         |         |                         |     |
|  -3 + *******.   +         +         +                         +     |
|     | ******.    |         |         |                         |     |
|  -4 + ***.       +         +         +                         +     |
|     | ****       |         |         |                         |  2  |
|  -5 + **.        +         +         +                         +     |
|     | **.        |         |         |                         |     |
|  -6 + ****.      +         +         +                         +     |
|     | ***.       |         |         |                         | --- |
|  -7 + .          +         +         +                         +     |
|     | ***.       |         |         |                         |     |
|  -8 + .          +         +         +                         +     |
|     |            |         |         |                         |     |
|  -9 + *          +         +         +                         + (1) |
|-----+------------+---------+---------+-------------------------+-----|
|Measr| * = 5      |-Rater   |-Session |-Item                    |Scale|
+-----------------------------------------------------------------------+
```

Figure 15. FACETS summary for the Tracking Essays. Each * equals 2 persons. Each . equals 1 person. Measr = Measure.

CHAPTER 5

RESULTS

This chapter presents the results of the quantitative and qualitative analyses in order to answer the hypotheses presented in Chapter 2. Each hypothesis was considered in turn, using data from the measures described in Chapter 3, as well as the qualitative data gathered from the participant interviews.

Hypothesis 1: Salience of the Reminders and Effect on Study Time

The first hypothesis was that the reminders would be more salient in the stronger treatments than the weaker treatments and the participants would, as a result, spend more time studying the target vocabulary. This hypothesis was explored using data acquired through the interviews conducted after the first semester in July and the three study instruments (Quizlet, Moodle, Vocabulary Self-study Logs).

Salience of the Reminders

In order to examine if the reminders had an effect on vocabulary learning, retention, and use, it must be confirmed that the reminders were salient to the participants. Table 21 shows a summary of the interviewees' responses indicating if they mentioned the reminders as part of the usual lesson proceedings. Eight of the 20 interviewees belonged to the comparison (NR) group, so they had not received the reminders. They were, therefore, excluded from Table 21.

Table 21. *Summary of Participant Responses Indicating Reminder Salience (n = 12)*

	Remind (R)	Remind + notice (R+)	Remind + use (R++)
Reminder mentioned	1	3	2
No mention	3	1	2

Six of the 12 interviewees did not mention the reminders, possibly due to a failure on my part to probe during the interviews. The absence of comments regarding the reminders does not necessarily mean that they did not notice them. Rather, my interview might not have elicited a full response. Out of the 12 interviewees who received reminders, five mentioned that using the target vocabulary was part of their homework. When asked to describe a typical lesson, these interviewees' responses included the following:

02-R++: My teacher, every time, he told us to use the words. So I could remember and try to use them when I write. (line 223)

03-R: I did my homework, I watched this printout. My CC1 teacher's homework is, use this column in the report. (line 241)

07-R+: So I remember to use, I have to use those words. (line 204)

09-R+: My teacher said, homework … in homework, you use this. But this, it is very difficult for me to use this, only this week's vocabulary. (line 187)

14-R++: I wrote sometimes the NGSL for homework. But sometimes it is useful these words, but it is difficult to use it. So not me, but all of my class, are difficult to find to use 5 words. (line 173)

Five interviewees reported that they referred to the vocabulary list while completing their homework assignments, but they did not mention the reminders during the interview. In contrast, the interviewees from the NR group all stated that they did not use the vocabulary list for the CC1 homework. Most of them stated that they used the target vocabulary "naturally" or "just from memory." These two findings are important as they suggest that the participants who were reminded to use the target vocabulary consciously used material given in one class for homework in another. Interestingly, one participant from the NR group (lines 223-229) reported that she did not refer to the list despite being aware that her friends in other classes were being pushed to use the vocabulary.

Interviewer: When you were doing the narrative or report, when you were writing it, did you check the NGSL paper or just from memory?

18-NR: Just from memory. My class doesn't have the NGSL box like the other class. So, these students' paper had how many words did you choose from NGSL. But we don't have.

Interviewer: Okay. Did you try to use the NGSL words anyway?

18-NR: Mm (nods). But just naturally.

One interviewee from the R group reported that she did not refer to the vocabulary list; she made no mention of the reminder, perhaps because the weakest reminder treatment only asked the students to use five out of 50 words in their homework in a given week. Unlike the R+ and R++ group participants, who had to underline potentially useful vocabulary or write those words above their final draft, it is possible some students from the R group did not feel the need to use the target vocabulary, in part because there were no consequences for failing to do so.

However, one interviewee from the R+ group reported that, while she noticed the reminders, she forgot to refer to the vocabulary list. This finding is unsurprising because the R+ participants were only provided with a truncated vocabulary list while writing the final drafts of the three essays which might explain the lack of reference to it. The lower cognitive demands of the R+ task (highlighting/underlining potentially useful vocabulary) were likely a factor in the participants forgetting about the list during the revision process. Furthermore, this finding indicates that study habits also played a role in the use of material provided for a class.

Overall, the reminders were salient across the different treatment groups given that some participants explicitly mentioned the reminders in the interviews. Others, while not mentioning the reminders, implied an effect, especially when compared to the NR group. However, it must be noted that precision in the interview questions might have teased out more informative responses.

Study Time

As a result of the reminders being salient to the participants, it was hypothesized that the participants would spend more time studying the target vocabulary. This hypothesis was explored by analyzing the amount of study time as reported by Moodle and Quizlet use, and the Vocabulary Self-study Logs using the Jonckheere-Terpstra trend test with follow-up pairwise comparisons. The Jonckheere-Terpstra trend test was chosen because it checks for a difference between the medians of the groups and incorporates information about whether the order of the groups is meaningful. In short, the Jonckheere-Terpstra trend test was chosen as hypothesis 1 stated that the participants in stronger treatments would spend more time studying vocabulary than those in weaker treatments. The hypothesized order for the Jonckheere-Terpstra test was from smallest to largest. The study time in the second semester, when reminders were not administered, were also analyzed to evaluate if any effect of reminders on study time continued after the treatment had ceased.

For the Vocabulary Self-Study Logs collected at the end of the first semester, the reported amount of time spent on vocabulary study was significantly affected by the reminders, T_{JT} = 4,381.50, z = -2.77, p = .006. Ascending order pairwise comparisons with adjusted p-values showed that there were significant differences in reported study time between the R++ group compared to the NR group (p = .001), and the R group (p = .001). However, there were no significant differences in reported study time between the R++ and R+ groups (p > .05). There were also no significant differences in reported study time between the R+ group compared to the NR group (p = 1.00), and the R group (p = .87). Finally, there was no significant difference in reported study time between the NR group and R group (p = 1.00). In short, the Jonckheere-Terpstra trend test revealed a significant trend in the data: contrary to the hypothesis, as participants received stronger reminders, the reported amount of time spent on vocabulary study decreased.

The analysis revealed that only the data from SSLog 1, the self-study log in semester 1 when the groups were reminded, was significant. The data from SSLog 2 in the second semester was not significant, which confirms the finding that the reminders affected the amount of vocabulary study time. This result suggests that reminders need to be regularly administered for study time to be affected. Once the reminders cease, the participants are likely to revert to previous study habits. Quizlet (2017) use and Moodle (2016) quiz scores were not significant in Semester 1. However, Moodle quiz scores in Semester 2 were significant despite none of the participants receiving reminders at that time, $T_{JT} = 5{,}864.50$, $z = 2.52$, $p = .012$. These results were not expected. The likely cause was that the data from Quizlet use, which was not required, was unreliable as many of the participants did not use it, which resulted in large gaps in the data from all groups. Furthermore, the vocabulary self-study logs were self-reported and entirely reliant on participants' memory and/or honesty, which further affected the accuracy and reliability of the data.

Hypothesis 2: Effect of Study Time on Vocabulary Learning and Retention

Hypothesis 2 was that participants in stronger treatment conditions would demonstrate improved vocabulary learning and retention compared to weaker treatment conditions due to having spent more time studying the target vocabulary. This hypothesis is central to the entire study and is explored through analysis of the participants' NGSLT-Receptive and NGSLT-Productive Rasch logits. In addition, this hypothesis was explored by analyzing the participants' use of the target vocabulary in their tracking essays and supported with the Vocabulary Inventory Survey data.

Identifying the target vocabulary used in the tracking essays, and assessing if the words were already known, learned, or retained necessitated the following three comparisons: (a) TE1-NGSL2000, (b) TE2-TE1, and (d) TE3-TE2. Comparing TE1 to the NGSL2000 revealed the words that were considered already known by the participants. Comparing TE2 to TE1 showed the target vocabulary used in TE2 that were absent in TE1, suggesting that these words were learned after TE1 was administered. Finally, comparing TE3 and TE2 showed the target vocabulary that were retained or unused. Words that were categorized as learned in the first semester (i.e., those used in TE2) and subsequently used again in the second semester (TE3) were considered retained. The learned words that were not used in TE3 were classified as unused. These unused words might not have been used due to factors such as the writing topic, depth of vocabulary knowledge, perceived risk of using the words, or they were simply forgotten.

The comparisons were made using the Text Lex Compare online tool. The analysis identified the vocabulary types of the NGSL 501-1000 that were used in the tracking essays. Vocabulary types, the different forms (spellings) of words, was chosen over vocabulary families as the identification method because the results made it simple to identify discreet words from the vocabulary list that were used in the essays.

Receptive and Productive Vocabulary Rasch Logits (NGSLT-Receptive) at Time 1, Time 2, and Time 3

The second hypothesis stated that participants who received stronger reminder treatments would improve vocabulary learning and retention more than those in weaker treatments with the comparison (NR) group showing the least improvement. This hypothesis was further explored by examining the changes in NGSLT-Receptive and NGSLT-Productive Rasch logits for the participants at Times 1, 2, and 3. The mean Rasch logits at Time 2 were expected to mirror the strength of the reminder treatments, that is, on a graph showing the groups' mean NGSLT-Receptive or NGSLT-Productive Rasch logits, the R++ group would show the steepest gradient from Time 1 to Time 2, and the NR group would show the shallowest gradient in the same time period. Participants with missing scores were excluded from the analysis. I rated all tests (N = 172).

Table 22 shows the descriptive statistics for the NGSLT-Receptive Rasch logits for each of the groups at Times 1, 2, and 3. Figure 16 displays the progression of the mean Rasch logits for each of the groups across the data collection period. Both Table 22 and Figure 16 do not indicate that groups receiving stronger reminders performed better on the NGSLT-Receptive. However, the means for all of the groups increased from Time 1 to Time 2, and from Time 2 to Time 3.

Table 22. *Descriptive Statistics for the Groups' Receptive Vocabulary Rasch Logits at Times 1, 2, and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	2.11	1.10	72
	R	2.10	1.07	37
	R+	2.43	0.77	35
	R++	2.01	1.38	34
	Total	2.15	1.10	178
Time 2	NR	2.52	1.25	72
	R	2.48	0.85	37
	R+	2.60	0.84	35
	R++	2.33	1.32	34
	Total	2.49	1.11	178
Time 3	NR	2.70	0.95	72
	R	2.74	1.02	37
	R+	3.12	1.06	35
	R++	2.50	1.22	34
	Total	2.75	1.05	178

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group. Mean values are in Rasch logits.

Differences in Receptive Vocabulary Rasch Logits by Group. To evaluate if there was a significant difference between groups for type of reminder concerning the participants' receptive vocabulary (NGSLT-Receptive) Rasch logits, a mixed ANOVA was conducted. The within-subjects factor was time and the between-subjects factor was the strength of reminder. The dependent variable was the participants' NGSLT-Receptive Rasch logits.

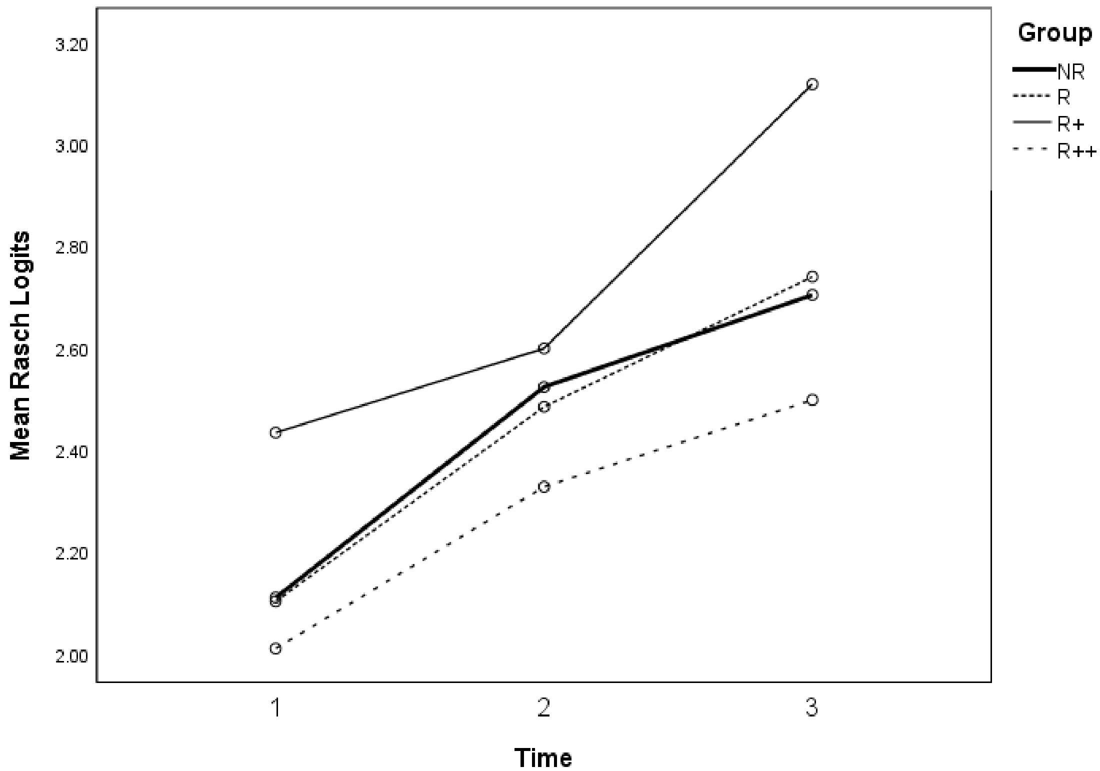

Figure 16. Receptive vocabulary Rasch logits for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

Five outliers were identified and removed from the analysis, as assessed by the examination of studentized residuals for values greater than ±3. The NGSLT-Receptive Rasch logits were normally distributed, as assessed by Normal Q-Q Plots. At Time 1, Levene's test indicated that the assumption of homogeneity of variances was not met ($p = .001$). However, the homogeneity of variances assumption was met ($p > .05$) at Times 2 and 3. The assumption of homogeneity of covariances was also not met, as assessed by

Box's test of equality of covariance matrices ($p = .024$). As a result, an interaction term was not interpreted for the NGSLT-Receptive.

The main effect of time showed a statistically significant difference in NGSLT-Receptive Rasch logits at the three time points, $F(2, 322) = 52.32$, $p < .001$, partial $\eta^2 = .245$. Follow-up pairwise comparisons were run where reported 95% confidence intervals and *p*-values were Bonferroni-adjusted. The unweighted marginal means of NGSLT-Receptive Rasch logits were 2.11 (SE = 0.09) at Time 1, 2.45 (SE = 0.08) at Time 2, and 2.72 (SE = 0.08) at Time 3. Mean NGSLT-Receptive Rasch logits increased from Time 1 to Time 2 by 0.34, 95% CI [0.20, 0.49], $p < .001$. Similarly, mean NGSLT-Receptive Rasch logits increased from Time 2 to Time 3 by 0.27, 95% CI [0.12, 0.42], $p < .001$. Overall, mean NGSLT-Receptive Rasch logits increased from Time 1 to Time 3 by 0.61, 95% CI [0.47, 0.76], $p < .001$.

The main effect of strength of reminder showed a non-significant difference in mean NGSLT-Receptive Rasch logits between treatment groups, $F(3, 161) = 1.75$, $p = .16$, partial $\eta^2 = .03$.

Differences in Productive Vocabulary Rasch Logits by Group. Unlike the receptive test (NGSLT-Receptive), the productive test was expected to be more difficult. Therefore, changes in NGSLT-Productive Rasch logits were expected to be to a lesser degree than the NGSLT-Receptive. Participants with missing scores were excluded from the analysis. All tests were rated by the researcher. Table 23 shows the descriptive statistics for the NGSLT-Productive Rasch logits for each of the groups at Times 1, 2, and 3. Figure 17 displays the progression of the mean Rasch logits for each of the groups across the data collection period.

Table 23. *Descriptive Statistics for the Groups' Productive Vocabulary Rasch Logits at Times 1, 2, and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	-0.13	1.01	72
	R	-0.03	0.79	37
	R+	0.24	0.98	35
	R++	-0.10	1.20	34
	Total	-0.03	1.00	178
Time 2	NR	1.12	1.43	71
	R	0.88	0.72	37
	R+	1.50	1.02	34
	R++	0.87	1.33	33
	Total	1.10	1.23	175
Time 3	NR	1.28	1.16	71
	R	1.24	0.80	36
	R+	1.57	1.00	33
	R++	1.17	1.54	32
	Total	1.31	1.15	172

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group. Mean values are in Rasch logits.

Both Table 23 and Figure 17 indicate that groups receiving stronger reminders did not perform better on the NGSLT-Productive test from Time 1 to Time 2; the NR group showed a greater rate of improvement than the treatment groups.

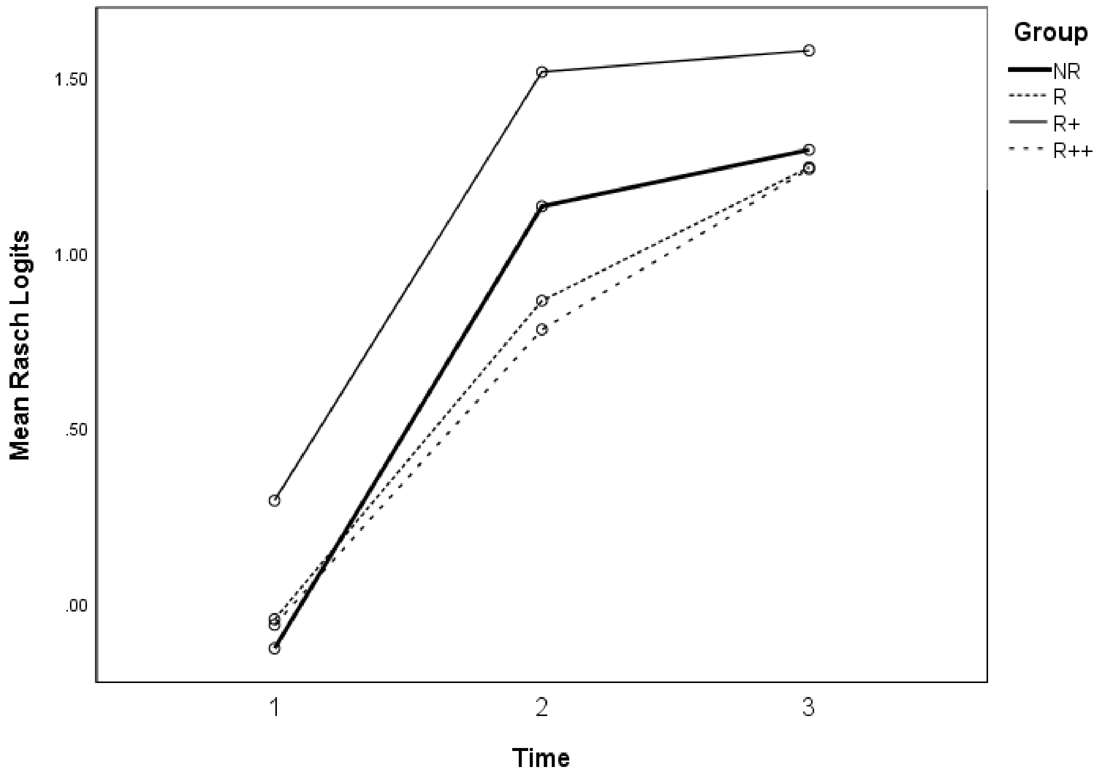

Figure 17. Mean productive vocabulary Rasch logits for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

Eight outliers were removed from the analysis, as assessed by an examination of studentized residuals for values greater than ±3. The NGSLT-Productive Rasch logits were normally distributed, as assessed by Normal Q-Q Plots. At Time 1, the homogeneity of variances assumption was met ($p > .05$) as assessed by Levene's test. However, the assumption was not met at Times 2 and 3 ($p < .05$). The assumption of homogeneity of covariances was met ($p = .007$) as assessed by Box's M test. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the two-way interaction, $\chi^2(2)$

= 15.57, $p < .001$; therefore, Huynh-Feldt corrected degrees of freedom are reported ($\varepsilon = .94$). There was no statistically significant interaction between the intervention and time on NGSLT-Productive Rasch logits, $F(5.65, 297.50) = 1.47$, $p = .19$, partial $\eta^2 = .03$; therefore, analyses of the main effects for time and group were performed.

The main effect of time showed a statistically significant difference in NGSLT-Productive Rasch logits at the three time points, $F(2, 316) = 265.70$, $p < .001$, partial $\eta^2 = .63$. Follow-up pairwise comparisons were run where reported 95% confidence intervals and *p*-values were Bonferroni-adjusted. The unweighted marginal means of NGSLT-Productive Rasch logits were -0.07 (SE = 0.07) at Time 1, 0.95 (SE = 0.09) at Time 2, and 1.22 (SE = 0.08) at Time 3. Mean NGSLT-Productive Rasch logits increased from Time 1 to Time 2 by 1.02, 95% CI [0.86, 1.18], $p < .001$. Similarly, mean NGSLT-Productive Rasch logits increased from Time 2 to Time 3 by 0.27, 95% CI [0.14, 0.39], $p < .001$. Overall, mean NGSLT-Productive Rasch logits increased from Time 1 to Time 3 by 1.29, 95% CI [1.14, 1.43], $p < .001$.

The main effect of strength of reminder showed that there was a non-significant difference in mean NGSLT-Productive Rasch logits between treatment groups, $F(3, 158) = 2.45$, $p = .07$, partial $\eta^2 = .044$.

Target Vocabulary in the Tracking Essays

Because it was hypothesized that the reminders would lead to more time spent studying the target vocabulary, it follows, then, that stronger reminder treatments would result in participants learning more vocabulary in the first semester, which would, in turn, lead to a greater number of target vocabulary usage in the tracking essays. It was also hypothesized that stronger reminder treatments would result in participants retaining more vocabulary in the second semester even after the reminders had ceased. Thus, groups receiving stronger reminder treatments were expected to show greater vocabulary learning and retention compared to those receiving weaker treatments with the comparison group showing the least gains. Table 24 reports a summary of the vocabulary known at the outset of the study, learned during the first semester, retained through the second semester, and unused in Tracking Essay 3 at the end of the data collection period. *Unused*, in the context of this study, refers to target vocabulary that was not used in Tracking Essay 1, used in Tracking Essay 2, but not used again in Tracking Essay 3. The numbers indicate the sum of target vocabulary in each category for each group.

Table 24. *Summary of Target Vocabulary Used in Tracking Essays*

	Known	Learned	Retained	Unused
NR (*n* = 72)	659	93	40	53
R (*n* = 37)	492	69	30	39
R+ (*n* = 35)	495	60	28	32
R++ (*n* = 34)	198	75	29	46

Note. Pretest = vocabulary from the 2,000-word NGSL that were used in TE1, Learned = vocabulary from the second 500 of the NGSL that was learned during term 1 and used in TE2 but not used in TE1, Retained = vocabulary from the second 500 of the NGSL that was learned in term 1 and used in TE3 and in TE2, Unused = vocabulary that was learned during term 1 and used in TE2 but not used in TE3, NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

This initial count of target in the four categories did not reveal any patterns between the groups that indicated whether the hypothesis was confirmed or not. In fact, the most noticeable difference was observed in the comparison (NR) group, which demonstrated a much larger amount of vocabulary learning compared to the groups that received the reminders as denoted by the *Learned* column. However, the relatively larger number of learned vocabulary can likely be attributed to the greater *n*-size of the group that would have the effect of contributing a larger variety of vocabulary in the analysis, thereby increasing the number of different target vocabulary items used in the tracking essays. While this first step did not contain the necessary detail in determining if greater reminders lead to greater vocabulary learning or retention, the results suggested that learning and retention of the target vocabulary occurred in all four groups to some extent.

It was important to estimate the amount of vocabulary known by the participants at the outset of the study. The descriptive statistics of the known vocabulary are displayed in Table 25. Prior to the treatments being administered, the mean number of known vocabulary of the groups showed no particular order in relation to strength of treatment with the R condition showing the lowest mean number of known target vocabulary (52.84), followed by the R++ condition (54.38), NR condition (55.14), and R+ condition (59.06).

Table 25. *Descriptive Statistics of the Participants' Known Vocabulary*

	NR (*n* = 71)	R (*n* = 37)	R+ (*n* = 34)	R++ (*n* = 34)
M	55.14	52.84	59.06	54.38
SE	1.88	2.58	2.91	3.61
95% CI	[51.38, 58.90]	[47.60, 58.07]	[53.14, 64.97]	[47.04, 61.73]
SD	15.87	15.70	16.95	21.06
Skewness	0.35	0.28	0.57	0.76
SES	0.29	0.39	0.40	0.40
z-Skewness	1.21	0.72	1.43	1.9
Kurtosis	0.88	-0.93	-0.16	0.15
SEK	0.56	0.76	0.79	0.79
z-Kurtosis	1.57	-1.22	-0.20	0.19

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group, CI = confidence interval, SES = standard error skewness, SEK = standard error kurtosis.

At the end of the first semester, the number of vocabulary items learned was estimated by comparing Tracking Essay 2, completed in July, with Tracking Essay 1, completed in April. Descriptive statistics of the learned vocabulary are displayed in Table 26. The descriptive statistics show that the R, R+, and R++ groups had learned more vocabulary than the NR group and generally suggests that groups receiving stronger reminders learned more words. Comparing Tracking Essay 2 to Tracking Essay 1 showed that the participants in the NR condition learned the fewest target vocabulary ($M = 6.34$) compared to the treatment groups. The R++ participants learned the most target vocabulary ($M = 7.79$). The participants in the R and R+ conditions had means between the NR and R++ conditions at 6.81 and 6.47, respectively.

While the results shown in Table 26 suggest that the treatments resulted in improved vocabulary learning compared to the NR condition, it was also hypothesized that those participants would also demonstrate improved retention. Descriptive statistics of the retained vocabulary are displayed in Table 27. The descriptive statistics show a substantial difference in the mean number of retained vocabulary items between the stronger treatments (R+, R++) and the weaker treatments (NR, R). The R+ and R++ conditions had higher mean numbers of retained vocabulary, 1.94 and 2.03, respectively, compared to the NR and R conditions, 1.56 and 1.54, respectively.

Table 26. *Descriptive Statistics of the Participants' Learned Vocabulary*

	NR (*n* = 71)	R (*n* = 37)	R+ (*n* = 34)	R++ (*n* = 34)
M	6.34	6.81	6.47	7.79
SE	0.35	0.60	0.54	0.75
95% CI	[5.64, 7.04]	[5.59, 8.03]	[5.37, 7.57]	[6.27, 9.32]
SD	2.97	3.66	3.15	4.38
Skewness	0.64	0.85	0.40	0.73
SES	0.29	0.39	0.40	0.40
z-Skewness	2.21	2.18	1	1.83
Kurtosis	0.11	1.29	-0.08	-0.13
SEK	0.56	0.76	0.79	0.79
z-Kurtosis	0.20	1.70	-0.10	-0.16

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group, CI = confidence interval, SES = standard error skewness, SEK = standard error kurtosis.

Table 27. *Descriptive Statistics of the Participants' Retained Vocabulary*

	NR (*n* = 71)	R (*n* = 37)	R+ (*n* = 34)	R++ (*n* = 34)
M	1.56	1.54	1.94	2.03
SE	0.17	0.29	0.35	0.31
95% CI	[1.23, 1.90]	[0.95, 2.13]	[1.23, 2.65]	[1.40, 2.66]
SD	1.41	1.76	2.03	1.80
Skewness	0.95	1.37	1.56	1.31
SES	0.29	0.39	0.40	0.40
z-Skewness	3.28	3.51	3.90	3.28
Kurtosis	0.62	0.94	3.22	2.25
SEK	0.56	0.76	0.79	0.79
z-Kurtosis	1.11	1.24	4.08	2.85

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group, CI = confidence interval, SES = standard error skewness, SEK = standard error kurtosis.

Given that institutional restrictions made it impossible to ensure that participants wrote about the same topic for each Tracking Essay, it was also important to estimate the number of vocabulary items used in Tracking Essay 2 but not Tracking Essay 3. This estimation would aid in interpretation of the results as vocabulary that was not used again in Tracking Essay 3 could be due to different writing topics, depth of knowledge, perceived risk, or were forgotten. Descriptive statistics of the unused vocabulary are displayed in Table 28. The mean number of unused vocabulary items indicates that stronger treatments led to a higher incidence of unused vocabulary. The R++ condition showed the highest mean number of unused vocabulary (7.06), followed by the R+ (6.82), R (6.65), and NR (6.42) conditions. As stated above, this result might be attributed to different essay topics which can lead to different types of vocabulary being used.

One-way ANOVAs were conducted to assess the amount of target vocabulary that was learned, retained, or unused. All the assumptions of the ANOVA were checked and met. For learned vocabulary, defined as using vocabulary in TE2 that was absent in TE2, there was a non-significant effect of reminders on learning of vocabulary, $F(3, 172) = 1.45$, $p = .23$. For retained vocabulary, defined as using the learned vocabulary in TE3, there was a non-significant effect of reminders on retention of vocabulary, $F(3, 172) = .91$, $p = .44$. For unused vocabulary, defined as words that were used in TE2 but not used in TE3, the assumption of homogeneity of variance was violated; therefore, the Welch F-ratio is reported. There was a non-significant effect of reminders on unused vocabulary, $F(3, 172) = .24$, $p = .87$.

Table 28. *Descriptive Statistics of the Participants' Unused Vocabulary*

	NR (n = 71)	R (n = 37)	R+ (n = 34)	R++ (n = 34)
M	6.42	6.65	6.82	7.06
SE	0.37	0.48	0.59	0.78
95% CI	[5.69, 7.15]	[5.67, 7.63]	[5.63, 8.02]	[5.48, 8.64]
SD	3.09	2.94	3.42	4.53
Skewness	0.93	-0.01	0.73	0.96
SES	0.29	0.39	0.40	0.40
z-Skewness	3.21	-0.03	1.83	2.40
Kurtosis	0.55	-0.09	-0.06	0.03
SEK	0.56	0.76	0.79	0.79
z-Kurtosis	0.98	-0.12	-0.08	0.04

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group, CI = confidence interval, SES = standard error skewness, SEK = standard error kurtosis.

Overall, the one-way ANOVAs indicated that there were no significant differences between the groups. Despite the non-significant findings, the mean number of learned, retained, and unused vocabulary followed a linear trend that parallels the strength of reminders. However, the mean differences, as Figure 18 shows, is less than one item.

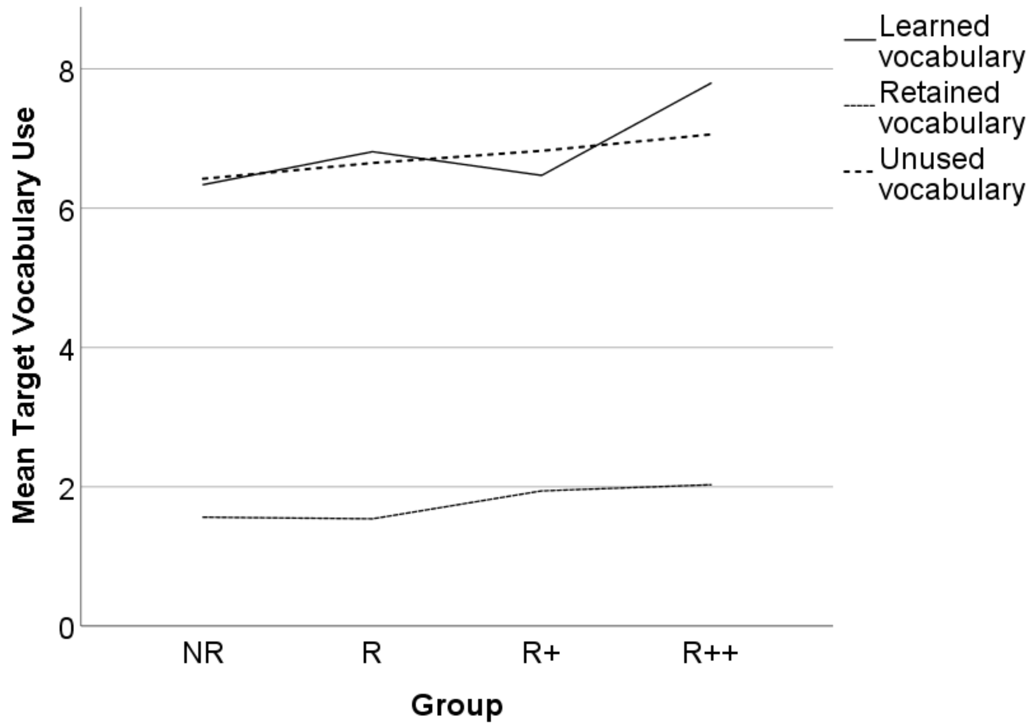

Figure 18. Mean target vocabulary learned, retained, and unused in Tracking Essay 2 for the groups; NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

In summary, the mean number of learned and retained vocabulary, as displayed in Table 26 and Table 27, indicate that the groups that received stronger reminders learned more vocabulary. However, these results were not statistically significant. Therefore, the second hypothesis was not confirmed. This trend continued in the reports of unused vocabulary (see Table 28). The results suggest that different topics led to a greater amount of unused vocabulary.

Correlation of Study Time With Vocabulary Test Scores

The second hypothesis also included the expectation that vocabulary study time would correlate positively with the vocabulary test scores. Correlation analyses were conducted between the following measures that were recorded over the first semester (i.e. between Time 1 and Time 2): Moodle (2016) use data, Quizlet (2017) use data, student self-study logs, and the receptive vocabulary test (NGSLT-Receptive) and productive vocabulary test (NGSLT-Productive) administered at Time 1 and Time 2. Supporting data came from the Vocabulary Inventory Surveys.

Correlation coefficients were computed among the three self-study measures and vocabulary test scale for Time 2 and Time 3. Time 1 was excluded from the analysis as the participants had not begun studying vocabulary at that time. Students with missing data were excluded from the analysis. Using the Bonferroni approach to control for Type I error across the ten correlations, a p value of less than .005 (.05/10 = .005) was required for significance.

The results of the correlational analysis for Time 2, presented in Table 29, show that only the amount of self-study using Moodle (2016) had a significant correlation with the observed vocabulary test scores. However, the strengths of association were small. The significant correlations between Moodle and NGSLT-Receptive mean that the two measures share 7.8% variance ($r^2 = 0.28^2 = 0.078$), which indicates that 7.8% of the variability in NGSLT-Receptive performance is shared by Moodle with 92.2% of the variability accounted for by other variables. Similarly, Moodle and the NGSLT-Productive share 4.4% variance ($r^2 = 0.21^2 = 0.044$). These results suggest that Moodle has little predictive power for performance on the vocabulary tests.

Table 29. *Correlations Among the Self-Study Measures and Total Vocabulary Test Score at Time 2 (N = 176)*

	Quizlet	Moodle	SSLog 1	NGSLT-Receptive
Moodle	.17	—	—	—
SSLog	.02	.07	—	—
NGSLT-Receptive	.03	.28*	.00	—
NGSLT-Productive	.13	.21*	.00	.60*

*$p < .005$

Note. SSLog 1 = student self-study log between Time 1 and Time 2, NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

The results of the correlational analysis for Time 3, presented in Table 30, show that the amount of self-study using Moodle (2016) had a significant correlation with the amount of independent vocabulary self-study as reported in the participants' self-study logs. The significant correlations between Moodle and the self-study logs means that the two measures share 9% variance ($r^2 = 0.30^2 = 0.09$). This result shows that the two study measures are not closely associated and suggests that Moodle use is not a predictor of reported vocabulary self-study nor reported self-study a predictor of Moodle use for vocabulary study.

Table 30. *Correlations Among the Self-Study Measures and Total Vocabulary Test Score at Time 3 (N = 174)*

	Quizlet	Moodle	SSLog 2	NGSLT-Receptive
Moodle	.11	—	—	—
SSLog	.09	.30*	—	—
NGSLT-Receptive	-.03	-.20	.10	—
NGSLT-Productive	-.01	-.09	.21	.65*

*p < .005

Note. SSLog 2 = student self-study log between Time 2 and Time 3, NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

Moodle use was the only estimate of study time that had a positive significant correlation with the vocabulary test scores at Time 2 and Time 3. This finding can likely be attributed to Moodle being part of the participants' course grade for Vocabulary Course 1 and Vocabulary Course 2. This meant that, while there was no penalty for not using Quizlet, there was a perceived risk for not completing the Moodle quizzes.

The data from the Vocabulary Inventory Surveys was analyzed using the Kruskal-Wallis test with a follow-up Jonckheere-Terpstra trend test to determine if there was a significant change in the reported amount of vocabulary known from Time 1 to Time 2 between the groups. The Kruskal-Wallis test was chosen as preliminary checks revealed that the assumptions of normality and homogeneity of variance were not met thereby making a one-way ANOVA inappropriate. The Kruskal-Wallis test indicated that there was not a significant change in reported amount of known vocabulary from Time 1 to Time 2 between the four groups, $H(3) = 5.219$, $p = .156$. The Jonckheere-Terpstra trend test did not reveal a significant trend in the data, $J = 5{,}147$, $z = -.752$, $p = .452$. These findings suggest that the strength of reminders did not affect vocabulary learning as evaluated by the Vocabulary Inventory Survey.

Correlations of Study Time With Target Vocabulary Use at Time 2

To further evaluate the second hypothesis, vocabulary study time was also expected to correlate positively with the amount of target vocabulary use. Correlation analyses were conducted between the following measures: Moodle (2016) use data, Quizlet (2017) use data, student self-study logs, and target vocabulary use data at Time 2. Target vocabulary use was defined as the number of target vocabulary that was used correctly or partially correctly in Tracking Essay 2 but was not used in Tracking Essay 1. Partially correct vocabulary use was included in the analysis as, being under test conditions, even partially correct use showed that the participant had some facility with the target vocabulary. There was no analysis of this at Time 1 because it was before reminders were given or any vocabulary study occurred. Similarly, an analysis at Time 3 was not conducted as the reminders were not given to any group. Using the Bonferroni approach to control for Type I error across the 6 correlations, a p value of less than .008 (.05/6 = .008) was required for significance (see Table 31).

Table 31. *Correlations Among the Self-Study Measures and Target Vocabulary Use at Time 2 (N = 176)*

	Quizlet	Moodle	SSLog 1
Moodle	.17	—	—
SSLog 1	.02	.07	—
Vocab. Use	-.03	.14	.07

*$p < .008$

Note. SSLog 1 = student self-study log between Time 1 and Time 2, Vocab. Use = target vocabulary used correctly and partially correctly in Tracking Essay 2 at Time 2.

The results of the correlational analysis presented in Table 31 show that the three self-study measures did not have the required level of significance of $p < .008$ for a correlation with target vocabulary use in Tracking Essay 2. This finding is surprising because in the previous correlational analyses, Moodle (2016) use correlated significantly with the vocabulary test scores at both Time 2 and Time 3.

A follow-up correlational analysis was conducted between the participants' NGSLT-Receptive and NGSLT-Productive scores at Time 2 and target vocabulary use. Students with missing data were excluded from the analysis. Using the Bonferroni approach to control for Type I error across the 2 correlations, a p value of less than .025 (.05/2 = .025) was required for significance (see Table 32).

The results of the correlational analysis of the NGSLT-Receptive, NGSLT-Productive, and target vocabulary used correctly or partially correctly show that all three variables were significantly correlated ($p < .025$). The significant correlation between the NGSLT-Receptive ($r = 0.20$) and NGSLT-Productive ($r = 0.19$) with target vocabulary use means that the target vocabulary use shares 4% and 3.6% of the variance with the NGSLT-Receptive and NGSLT-Productive, respectively. This finding is surprising as it indicates that vocabulary knowledge, as estimated by NGSLT-Receptive and NGSLT-Productive scores, has little predictive power for target vocabulary use in the writing task.

Table 32. *Correlations Between the Vocabulary Tests and Target Vocabulary Use at Time 2 (N = 176)*

	NGSLT-Receptive	NGSLT-Productive
NGSLT-Productive	.56*	—
Vocab. Use	.20*	.19*

**p* < .025

Note. SSLog 1 = student self-study log between Time 1 and Time 2, Vocab. Use = target vocabulary used correctly and partially correctly in Tracking Essay 2 at Time 2. NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

Hypothesis 3: Effect of Vocabulary Learning and Retention on GTEC and Writing Scores

The third hypothesis was that improved vocabulary learning and retention would lead to better performance on the GTEC and tracking essays. This hypothesis was explored by conducting a correlation analysis between the GTEC, NGSLT-Receptive, NGSLT-Productive, and FACETS fair averages of the tracking essays at Time 1, Time 2, and Time 3. A mixed ANOVA was also conducted to examine if there were significant differences between groups for type of reminder.

Correlation of the Vocabulary Scores with the GTEC and Writing Scores

Correlation coefficients were computed among the GTEC, NGSLT-Receptive, NGSLT-Productive, and tracking essay scores for each time period. Rasch logits for the NGSLT-Receptive and NGSLT-Productive, and FACETS fair average measures of the tracking essays were used. Using a Bonferroni adjustment to control for Type I error across the six correlations, a *p*-value of less than .008 (.05/6 = .008) was required for significance.

The results of the correlational analysis for Time 1 are presented in Table 33. Students with missing data were excluded from the analysis. All six correlations were statistically significant and were greater than or equal to .40. The positive correlation between the NGSLT-Receptive and NGSLT-Productive at Time 1 (r = .59) was expected because they are both tests of vocabulary. Similarly, the positive correlations the GTEC with the NGSLT-Receptive and NGSLT-Productive, r = .64 and r = 53, respectively, were expected because the GTEC is an English proficiency test and vocabulary is one aspect of overall language proficiency. Examination of the correlation coefficients for Tracking Essay 1 with the GTEC (r = .56), NGSLT-Receptive (r = .44), and NGSLT-Productive (r = .40), also made sense as overall language proficiency was expected to be closely related to writing ability, whereas vocabulary knowledge is only one component of language proficiency and writing ability. The great correlation of the GTEC with the NGSLT-

Receptive than the NGSLT-Productive was also expected as the GTEC is a receptive test of language proficiency.

Table 33. *Correlations Among the Four Test Scales at Time 1 (N = 178)*

	GTEC	NGSLT-Receptive	NGSLT-Productive
NGSLT-Receptive	.64	—	—
NGSLT-Productive	.53	.59	—
Tracking Essay 1	.56	.44	.40

All coefficients are significant at $p < .001$.
Note. GTEC = Global Test of English Communication, NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

At Time 2, the results of the correlational analysis presented in Table 34 show that all 3 of the correlations were statistically significant and were greater than or equal to .41. The positive correlations between the NGSLT-Receptive and NGSLT-Productive and between the Tracking Essay with the two vocabulary tests are similar to those that at Time 1 (see Table 33) and was expected.

Table 34. *Correlations Among the Four Test Scales at Time 2 (N = 178)*

	NGSLT-Receptive	NGSLT-Productive
NGSLT-Productive	.60	—
Tracking Essay 2	.39	.41

All coefficients are significant at $p < .001$.
Note. NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

At Time 3, the results of the correlational analysis presented in Table 35 show that all 6 correlations were statistically significant and were greater than or equal to .46. The correlation coefficients presented in Table 35 are similar to Table 33 although the correlation between NGSLT-Productive with the other measures is relatively stronger, suggesting that the participants' depth of vocabulary knowledge is improving.

Table 35. *Correlations Among the Four Test Scales at Time 3 (N = 178)*

	GTEC	NGLST-Receptive	NGSLT-Productive
NGLST-Receptive	.59	—	—
NGSLT-Productive	.59	.65	—
Tracking Essay 3	.65	.46	.46

All coefficients are significant at $p < .001$.
Note. GTEC = Global Test of English Communication, NGSLT-Receptive = New General Service List Test-Receptive, NGSLT-Productive = New General Service List Test-Productive.

GTEC Scores at Time 1 and Time 3

The descriptive statistics of the participants' GTEC scores are shown in Table 36. Figure 19 displays the participants' GTEC scores at Time 1 and Time 3. The results show that GTEC scores for all groups increased through Time 1 and Time 3.

The GTEC is designed to measure test-takers' listening and reading ability. All groups showed gains between Time 1 and Time 3 and there was no apparent indication that the reminders affected the rate of gains. The gains of all groups can be attributed to the English courses the participants were enrolled in.

Table 36. *Descriptive Statistics for the Groups' GTEC Scores at Times 1 and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	203.82	37.45	72
	R	206.11	25.18	37
	R+	207.97	22.91	34
	R++	198.62	48.00	34
	Total	204.10	35.15	178
Time 3	NR	216.26	29.67	72
	R	215.70	26.12	37
	R+	222.71	24.17	34
	R++	215.00	38.53	34
	Total	217.14	29.84	177

Note. GTEC = Global Test of English Communication, NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

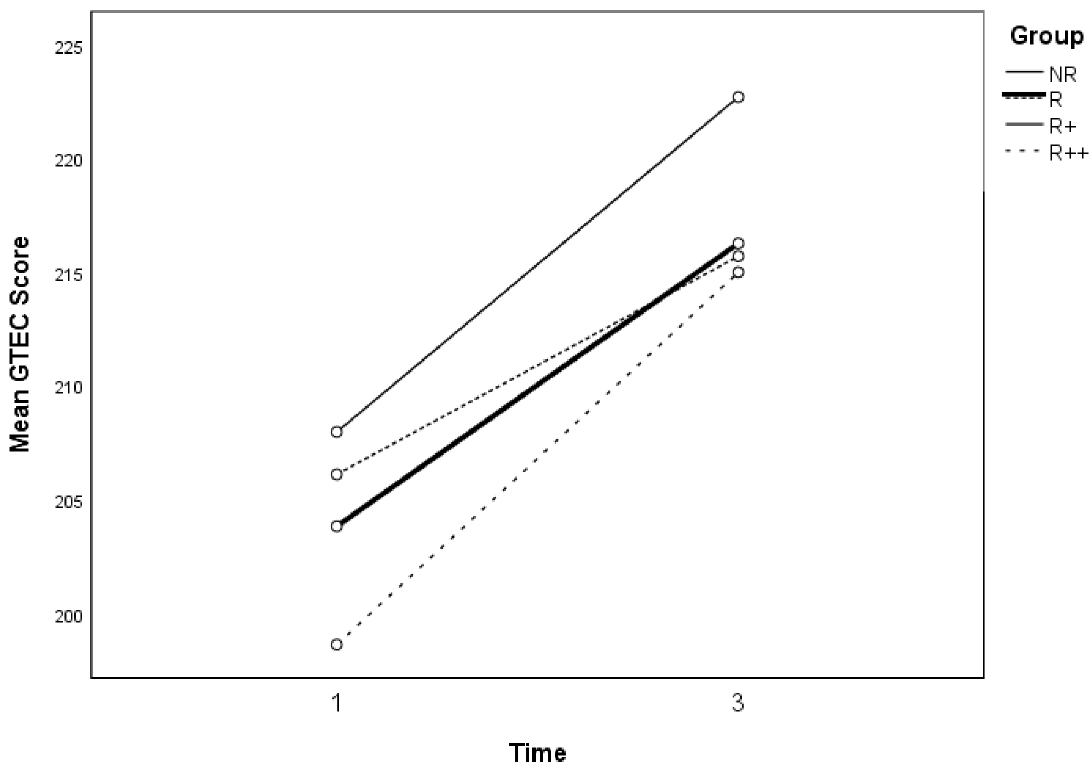

Figure 19. Mean GTEC scores for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

Differences in GTEC Scores by Group. To determine if there was a significant difference over time between groups for strength of reminder concerning the participants' GTEC scores at Time 1 and Time 3, a mixed ANOVA was conducted. The within-subjects factor was time and the between-subjects factor was the strength of reminder. The dependent variable was the participants' GTEC scores.

There were no outliers, as assessed by examination of studentized residuals for values greater than ±3. The GTEC scores were normally distributed, as assessed by Normal Q-Q Plots. The assumptions homogeneity of variances (p < .05) and covariances ($p < .05$) as assessed by Levene's test and Box's M test, respectively, were not met. As a result, an interaction term was not interpreted for the GTEC. For the GTEC, there are only two categories (i.e., Time 1 and Time 3), so the assumption of sphericity was not assessed. There was no statistically significant interaction between the intervention and time on GTEC scores, $F(3, 173) = 0.65$, $p = .58$, partial $\eta^2 = .011$. Therefore, analyses of the main effects for time and group were performed.

The main effect of time showed a statistically significant difference in GTEC scores at the two time points, $F(1, 173) = 58.99$, $p < .001$, partial $\eta^2 = .25$. Follow-up pairwise comparisons were run where 95% confidence intervals and *p*-values are Bonferroni-adjusted. The unweighted marginal means for GTEC scores were 204.13 (SE

= 2.79) at Time 1 and 217.42 (SE = 2.37) at Time 3. Mean GTEC scores increased from Time 1 to Time 3 by 13.29, 95% CI [9.87, 16.70], $p < .001$.

The main effect for strength of reminder showed that there was a non-significant difference in mean GTEC scores between treatment groups, $F(3, 173) = 0.45$, $p = .72$, partial $\eta^2 = .01$.

Hypothesis 4: Effect of Vocabulary Learning and Retention on Accurate Target Vocabulary Use in Writing

The fourth hypothesis was that improved vocabulary learning and retention for the reminder groups will enable the participants to use the target vocabulary more accurately in the tracking essays. This hypothesis was explored with the chi-square test for independence for Tracking Essay 2. The target vocabulary used in Tracking Essay 2 were coded as being correct, partially correct, or incorrect. Correct use of a word contained no grammatical errors in an embedded clause but could have minor spelling errors (phonetically similar to the correct spelling). Examples of correct usage include, *the express <u>bus</u>*, *we can go to the <u>traditional</u> shrine*, and *I <u>danced</u> on the stage*. Partially correct usage occurred with minor grammatical errors that did not impede understanding of the sentence. Examples of partially correct usage include, *We could not <u>prepare</u> the dance well*, *Then ride rollercoaster for <u>everyone</u>*, and *The <u>travel</u> for the foreign country is*

2 times for me. Incorrect use was where the word was used inappropriately in terms of collocational use, contextual fit, or meaning. Examples of incorrect usage included *I think that the* <u>*conclusion*</u> *became better if I do my best through these experience* and using the target word as a proper noun such as in *Hollywood* <u>*Dream*</u> *the Ride*.

All target vocabulary used was rated by me and 20 (11.36%) participants were randomly selected to be evaluated by a second rater to establish inter-rater reliability. In total, 151 words were rated by the second rater. Cohen's kappa was run to evaluate the degree of agreement between the two raters' judgments of whether the 151 words were correctly, partially correctly, or incorrectly used. The analysis indicated a substantial strength of agreement between the two raters' judgements, $\kappa = .71$ (95% CI, .59 to .82), $p < .01$. Table 37 displays the degree to which the two raters agreed on the correctness of target vocabulary use.

Table 37. *Crosstabulation Table of Rated Correctness by Two Raters*

		Rater 1			
		Correct	Partially correct	Incorrect	Total
Rater 2	Correct	101	10	1	112
	Partially Correct	3	27	0	30
	Incorrect	1	4	4	9
Total		105	41	5	151

The contingency table shown in Table 38 displays the number of learned target vocabulary used correctly, partially correctly, and incorrectly for each group. The percentage of identified instances within groups (% within Group) is an indicator of a treatment effect. The NR group showed correct target vocabulary use in 66.15% of the identified instances. The R group showed 69.38%, the R+ group showed 70.22%, and the R++ group showed 69.43%. As seen in Table 38, the NR group produced 299 correct vocabulary items, which was lower than the expected count of 308.90. In contrast, the three reminder groups all had higher numbers of correct items (R = 179, R expected = 176.30; R+ = 158, R+ expected = 153.75; R++ = 184, R++ expected = 181.08).

The results also showed that target vocabulary use was rated as partially correct in 28.54% of the identified instances for the NR group (see Table 38). The R group (27.52%), R+ group (24.89%), and R++ group (25.28%) had lower percentages of correct usage. Similarly, vocabulary use was rated as incorrect in 5.31% of instances for the NR group, 3.10% for the R group, 4.89% for the R+ group, and 5.28% for the R++ group. All reminder groups had higher percentages of correct target vocabulary use and lower percentages of partially correct use compared to the comparison (NR) group. The results of the chi-square test for independence showed that the groups receiving stronger reminders had a higher percent of correct and partially correct use of the target vocabulary. In summary, there was a non-significant association between the types of

Table 38. *Contingency Table of Rated Correctness in Vocabulary Use*

Group		Vocabulary use rating: Correct	Partially correct	Incorrect	Total
NR	Count	299_a	129_a	24_a	452
n = 71	Expected count	308.87	121.66	21.47	452
	% within group	66.15%	28.54%	5.31%	100.00%
	% within vocabulary use	36.46%	39.94%	42.11%	37.67%
	% of total	24.92%	10.75%	2.00%	37.67%
	Standardized residual	-0.56	0.67	0.55	
R	Count	179_a	71_a	8_a	258
n = 37	Expected count	176.30	69.45	12.26	258
	% within group	69.38%	27.52%	3.10%	100.00%
	% within vocabulary use	21.83%	21.98%	14.04%	21.50%
	% of total	14.92%	5.92%	0.67%	21.50%
	Standardized residual	0.20	0.19	-1.22	
R+	Count	158_a	56_a	11_a	225
n = 34	Expected count	153.75	60.56	10.69	225
	% within group	70.22%	24.89%	4.89%	100.00%
	% within vocabulary use	19.27%	17.34%	19.30%	18.75%
	% of total	13.17%	4.67%	0.92%	18.75%
	Standardized residual	0.34	-0.59	0.10	
R++	Count	184_a	67_a	14_a	265
n = 33	Expected count	181.08	71.33	12.59	265
	% within group	69.43%	25.28%	5.28%	100.00%
	% within vocabulary use	22.44%	20.74%	24.56%	22.08%
	% of total	15.33%	5.58%	1.17%	22.08%
	Standardized residual	0.22	-0.51	0.40	
Total	Count	820	323	57	1200
N = 175	Expected count	820.00	323.00	57.00	1200
	% within group	68.33%	26.92%	4.75%	100.00%
	% within vocabulary use	100.00%	100.00%	100.00%	100.00%
	% of total	68.33%	26.92%	4.75%	100.00%

Note. Each subscript letter denotes a subset of vocabulary use categories whose column proportions do not differ significantly from each other at the .05 level. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

reminders and amount of target vocabulary used correctly, partially correctly, or incorrectly, χ^2 (1) = 3.55, p = .74 suggesting that the reminders did not have a significant effect on correctness of vocabulary use.

Relationship Between Vocabulary Use and Writing Scores

The fourth hypothesis also stated that writing scores would be higher in stronger treatment conditions. This question was explored by: (a) examining the writing quality of the tracking essays by using the JACET writing rubric with follow-up analyses using FACETS, (b) analyzing the lexical complexity of the tracking essays by using the formula WT^2/W (word types squared divided by the total number of words) (Ong & Zhang, 2010), and (c) calculating the number of standard length words for TE1, TE2 (treatment and comparison groups), and TE3 (treatment and comparison groups). The participants completed Tracking Essay 1 in early April (Time 1), Tracking Essay 2 in mid-July (Time 2), and Tracking Essay 3 in mid-January (Time 3). The participants wrote each essay under test conditions during a 30-minute period.

JACET writing rubric scores. It was hypothesized that stronger reminder treatments would result in participants making significant increases in their Tracking Essay fair averages due, in part, to greater usage of contextually appropriate target vocabulary. This hypothesis means that groups receiving stronger reminder treatments were expected to show greater gains compared to those receiving weaker treatments with the comparison (NR) group showing the least improvement. This hypothesis was explored by examining the changes in the participants' Tracking Essay Rasch logit ability estimates from the three measurement periods. All essays (N = 531) were rated by the researcher and 55 essays (10.36%) were randomly selected to be rated by a second rater to evaluate Rasch rater fit. The procedure for randomly selecting the 55 co-rated essays consisted of the following steps: (a) All the essays were first sorted by the five Content Course 1 (CC1) levels followed by the participants' last names, and (b) a random integer set generator (Random.org, 2019) was used to randomly select 11 tracking essays from each of the five levels. The raters applied the JACET writing rubric (Kuru et al., 2009), which consists of a five-point scale (1 to 5) for five criteria: content & idea development, organization, grammar, vocabulary, and mechanics (see Chapter 3 above). Following the rating, the fair averages of the participants were calculated using FACETS.

The results of the FACETS analysis indicated that the raters did not vary considerably in severity with rater 2 at .20 logits, and rater 1 at -.20 logits. This means that rater 2 was more severe than rater 1. Both raters met the infit and outfit mean square criteria of 0.50 to 1.50. Both raters had standardized statistics within the ±2.00 criterion. The Rasch reliability index of .87 indicates the degree to which these severity measures would be reproduced if the raters evaluated a similar sample of essays. The reliability of .87 is fairly high because there is some variance in rater severity. This suggests a good level of rater agreement (according to Fischer, 2007, a person and item measurement reliability of .81 to .90 is considered good). For inter-rater reliability, the observed Rasch agreement (49.8%) and expected Rasch agreement (47.8%) indicated that the two raters showed acceptable fit to the Rasch model. Table 39 reports the rater severity and fit statistics for the two raters who provided ratings for 55 essays (10.36%).

The Session Measurement Report (see Table 18 in Chapter 4) shows that the fair average, the expected observed score if the candidate had been rated by an unbiased average severity judge, for the 55 co-rated essays increased from Time 1 to Time 2 and from Time 2 to Time 3.

Table 40 shows the descriptive statistics for the tracking essay fair averages for each of the groups at Times 1, 2, and 3. Figure 20 displays the progression of the mean Rasch logits for each of the groups across the data collection period. Both Table 40 and Table 40 shows the descriptive statistics for the tracking essay fair averages for each of the groups at Times 1, 2, and 3. Figure 20 displays the progression of the mean Rasch logits for each of the groups across the data collection period. Both Table 40 and Figure 17 do not indicate that groups receiving stronger reminders performed better on the

NGSLT-Receptive. However, the means for all four groups increased from Time 1 to Time 2, and from Time 2 to Time 3.

Table 39. *Rater Measurement Report*

	Rater 1	Rater 2	*M*	*SD* (Sample)
Total score	737	699	718	26.90
Total count	275	275	275	0.00
Obsvd Average	2.68	2.54	2.61	0.10
Fair Average (*M*)	2.65	2.50	2.58	0.00
Measure	-0.20	0.20	0.00	0.29
Model *SE*	.10	.10	.10	.00
Infit Mnsq	0.92	1.07	1.00	0.10
Infit Zstd	-0.90	0.80	0.00	1.30
Outfit Mnsq	0.92	1.09	1.01	0.12
Outfit Zstd	-0.90	1.10	0.10	1.40
Estim. Discrm	1.08	0.92		
Correlation PtMea	.77	.76	.77	.01
Correlation Pt Exp	.77	.77		
Exact Agree. Obs %	49.80	49.80		
Exact Agree. Exp %	47.80	47.80		

Note. Obsvd Average = observed average, Mnsq = mean-square fit values, Zstd = standardized statistics, Estim. Discrm = estimated discrimination, PtMea = point-measure correlations, Pt Exp = expected value, Exact Agree. Obs% = observed percentage of exact agreements between raters, Exact Agree. Exp% = expected percentage of exact agreements between raters on ratings.

Table 40. *Descriptive Statistics for the Groups' Tracking Essay Fair Averages at Times 1, 2, and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	2.44	0.83	70
	R	2.34	0.82	36
	R+	2.63	0.66	33
	R++	2.43	1.00	31
	Total	2.45	0.83	170
Time 2	NR	2.76	0.75	70
	R	2.72	0.73	36
	R+	2.91	0.73	33
	R++	2.77	0.82	31
	Total	2.78	0.75	170
Time 3	NR	2.95	0.78	70
	R	3.01	0.71	36
	R+	3.05	0.77	33
	R++	2.97	0.81	31
	Total	2.98	0.76	170

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group. Mean values are in Rasch fair averages.

Differences in Tracking Essay fair averages by group. To determine if there was a significant difference between groups for type of reminder concerning the participants' tracking essay fair averages, a mixed ANOVA was conducted. The within-subjects factor was time and the between-subjects factor was the strength of reminder. The dependent variable was the participants' tracking essay fair averages.

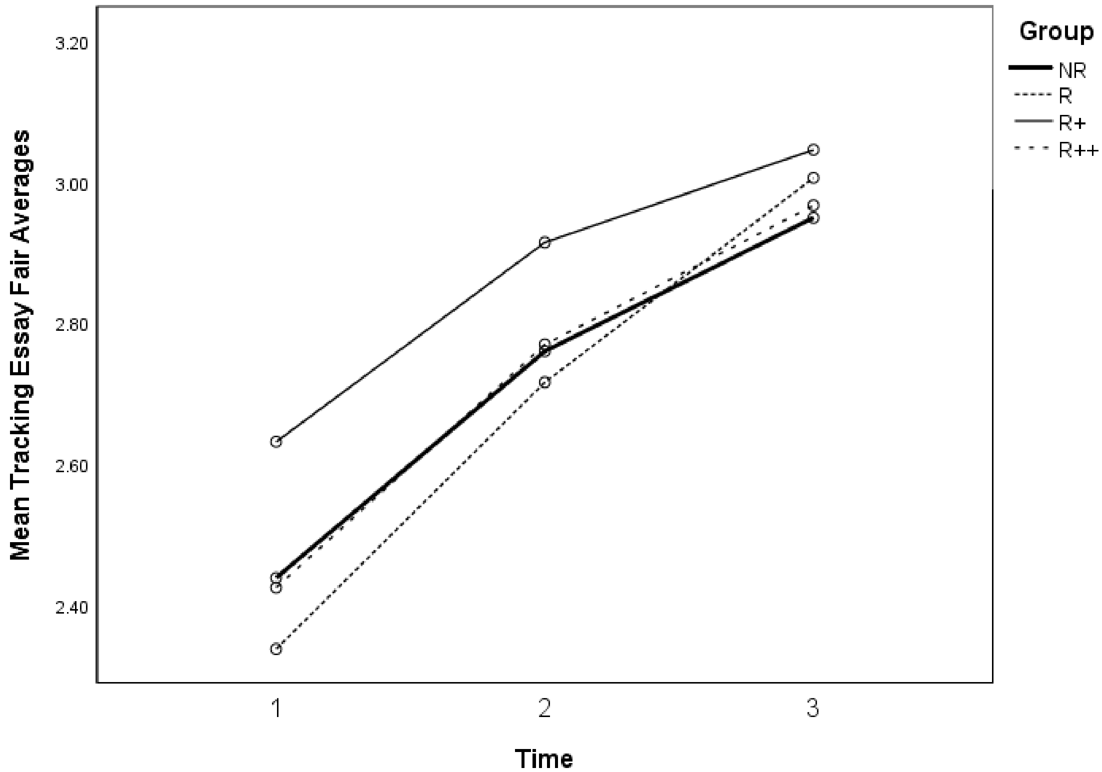

Figure 20. Mean Tracking Essay fair averages for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

There were no outliers, as assessed by an examination of studentized residuals for values greater than ±3. The tracking essay scores were normally distributed, as assessed by Normal Q-Q Plots. The homogeneity of variances assumption ($p > .05$) and the homogeneity of covariances assumption ($p = .22$) were met, as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the two-way interaction, $\chi^2(2) = 16.66$, $p < .001$; therefore Huynh-Feldt corrected degrees of freedom are reported (ε

= .98). There was no statistically significant interaction between the intervention and time on tracking essay fair averages, $F(5.63, 311.64) = 0.60$, $p = .72$, partial $\eta^2 = .01$. Therefore, analyses of the main effects for time and group were performed.

The main effect of time showed a statistically significant difference in tracking essay fair averages at the different time points, $F(1.88, 311.64) = 67.74$, $p < .001$, partial $\eta^2 = .29$. Follow-up pairwise comparisons were run where reported 95% confidence intervals and *p*-values were Bonferroni-adjusted. The unweighted marginal means of tracking essay fair averages were 2.46 (SE = 0.07) at Time 1, 2.79 (SE = 0.06) at Time 2, and 2.99 (SE = 0.06) at Time 3. Mean tracking essay fair averages increased significantly from Time 1 to Time 2 by 0.33, 95% CI [0.21, 0.45], $p < .001$. Similarly, mean tracking essay fair averages increased significantly from Time 2 to Time 3 by 0.20, 95% CI [0.11, 0.30], $p < .001$. Overall, mean tracking essay fair averages increased from Time 1 to Time 3 by 0.53, 95% CI [0.41, 0.66], $p < .001$.

The main effect of strength of reminder showed that there was a non-significant difference in mean tracking essay fair averages between treatment groups, $F(3, 166) = 0.43$, $p = .73$, partial $\eta^2 = .008$.

Lexical complexity of the Tracking Essays. As a component of writing scores, higher lexical complexity scores for participants in stronger treatments than those in weaker treatment groups was also a part of the fourth hypothesis. This means that participants who received stronger reminders would learn more vocabulary which would potentially result in them possessing a greater amount of productive vocabulary to draw from during the writing task. The lexical complexity of the tracking essays was calculated using the formula WT^2/W (word types squared divided by the total number of words) (Ong & Zhang, 2010). The number of words and word types needed for the calculation was obtained from the VocabProfiler online resource (Cobb, n.d.-b). Each participant's Tracking Essay was entered into VocabProfiler which reported the number of total words in the text (tokens) and the number of different words (types).

Table 41 shows the descriptive statistics for the lexical complexity for each of the groups at Times 1, 2, and 3. Figure 21 displays the progression of the mean lexical complexity for each of the groups across the data collection period. Both Table 41 and Figure 21 do not indicate that groups receiving stronger reminders had more lexical complexity. However, the means for all of the groups increased from Time 1 to Time 2, and from Time 2 to Time 3.

Table 41. *Descriptive Statistics for the Groups' Lexical Complexity at Times 1, 2, and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	44.84	11.15	70
	R	41.59	10.85	37
	R+	46.57	12.67	34
	R++	42.55	13.42	33
	Total	44.05	11.88	174
Time 2	NR	49.07	10.55	70
	R	51.00	12.51	37
	R+	52.93	10.97	34
	R++	48.53	13.05	33
	Total	50.13	11.58	174
Time 3	NR	53.95	13.28	30
	R	50.99	11.56	37
	R+	56.14	14.54	34
	R++	51.49	14.80	33
	Total	53.28	13.51	174

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

Differences in lexical complexity by group. To determine if there was a significant difference between groups for type of reminder concerning the participants' lexical complexity, a mixed ANOVA was conducted. The within-subjects factor was time and the between-subjects factor was the strength of reminder. The dependent variable was the participants' lexical complexity values.

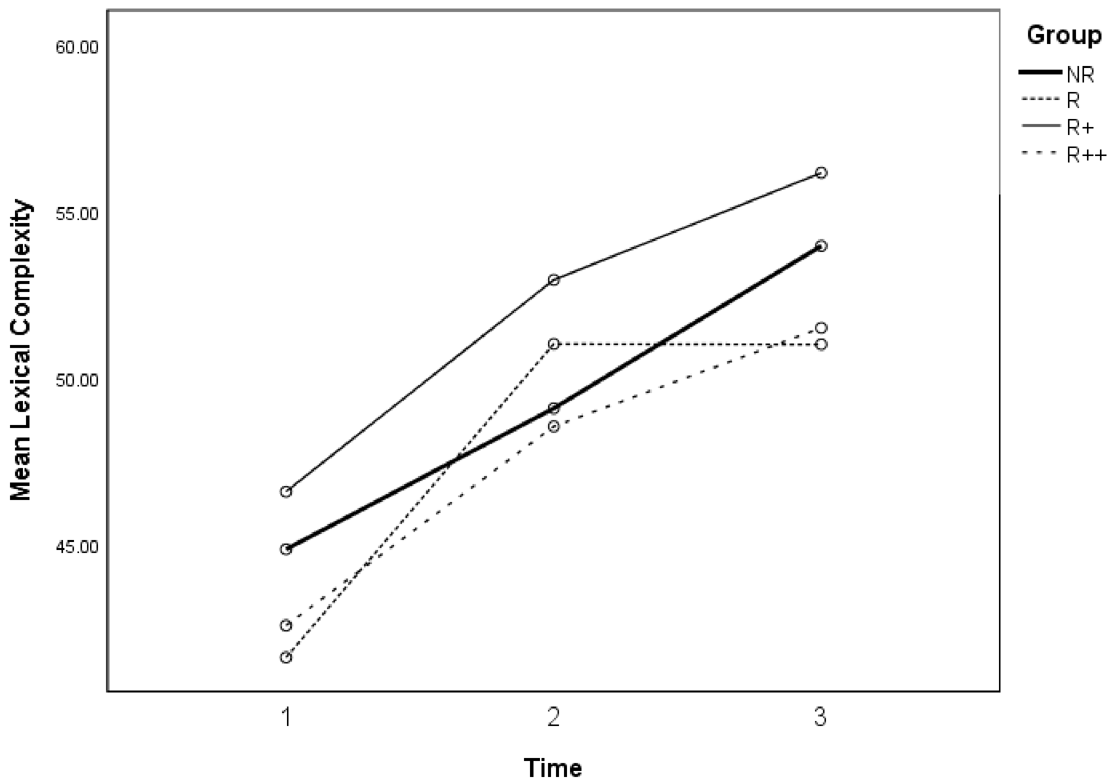

Figure 21. Mean lexical complexity for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

Two outliers were identified and removed from the analysis, as assessed by examination of studentized residuals for having values greater than ±3. The measures of lexical complexity were normally distributed, as assessed by Normal Q-Q Plots. The homogeneity of variances assumption ($p > .05$) and homogeneity of covariances assumption ($p = .90$) were met as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the two-way interaction, $\chi^2(2) = 9.67$, $p = .008$, therefore

Huynh-Feldt corrected degrees of freedom are reported ($\varepsilon = .98$). There was no statistically significant interaction between the intervention and time on lexical complexity, $F(5.85, 331.37) = 1.07$, $p = .38$, partial $\eta^2 = .02$. Therefore, analyses of the main effects for time and group were performed.

The main effect of time showed a statistically significant difference in lexical complexity at the different time points, $F(1.95, 331.37) = 55.55$, $p < .001$, partial $\eta^2 = .25$. Follow-up pairwise comparisons were run where reported 95% confidence intervals and *p*-values were Bonferroni-adjusted. The unweighted marginal means of lexical complexity were 43.89 (SE = 0.94) at Time 1, 50.38 (SE = 0.92) at Time 2, and 53.14 (SE = 1.07) at Time 3. Mean lexical complexity increased from Time 1 to Time 2 by 6.49, 95% CI [4.58, 8.41], $p < .001$, by 2.76, 95% CI [0.52, 5.00], $p = .01$, from Time 2 to Time 3, and by 9.25, 95% CI [6.90, 11.61], $p < .001$, from Time 1 to Time 3.

The main effect of strength of reminder showed that there was no significant difference in mean lexical complexity between treatment groups, $F(3, 170) = 1.23$, $p = .30$, partial $\eta^2 = .02$.

Number of standard length words of the Tracking Essays. As a component of writing scores, the participants in stronger treatments would demonstrate greater depth of vocabulary knowledge than those in weaker treatment groups. This would be reflected by greater ease of writing when completing the tracking essays, meaning that the participants in stronger treatments would write more over time, as measured by standard length words. However, this gain is also likely partly attributed to the participants having received essay-writing and other English-language instruction throughout the data-collection period. The number of standard length words for the tracking essays was determined by counting the number of character spaces, where a character space consists of an alphabetical character as well as a space between words and punctuation marks, and dividing this number by 6 (Carver, 1976). Microsoft Word reports a count of *Characters (with spaces)* of a text, which is equivalent to the number of character spaces.

Table 42 shows the descriptive statistics for the number of standard length words for each of the groups at Times 1, 2, and 3. Figure 22 displays the progression of the mean number of standard length words for each of the groups across the data collection period. Both Table 42 and Figure 22 do not indicate that groups receiving stronger reminders had more lexical complexity. However, the means for all of the groups increased from Time 1 to Time 2, and from Time 2 to Time 3.

Table 42. *Descriptive Statistics for the Groups' Number of Standard Length Words at Times 1, 2, and 3*

	Group	*M*	*SD*	*N*
Time 1	NR	151.32	55.79	70
	R	150.55	62.11	37
	R+	156.15	45.83	33
	R++	131.68	52.66	31
	Total	148.52	55.06	171
Time 2	NR	200.83	69.01	70
	R	219.56	66.14	37
	R+	213.05	51.60	33
	R++	192.54	67.71	31
	Total	205.74	65.29	171
Time 3	NR	231.48	74.56	70
	R	228.73	60.67	37
	R+	231.47	61.04	33
	R++	208.36	66.61	31
	Total	226.69	67.80	171

Note. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

Differences in number of standard length words by group. To determine if there was a significant difference between groups for type of reminder concerning the participants' number of standard length words, a mixed ANOVA was conducted. The within-subjects factor was time and the between-subjects factor was the strength of reminder. The dependent variable was the participants' number of standard length words.

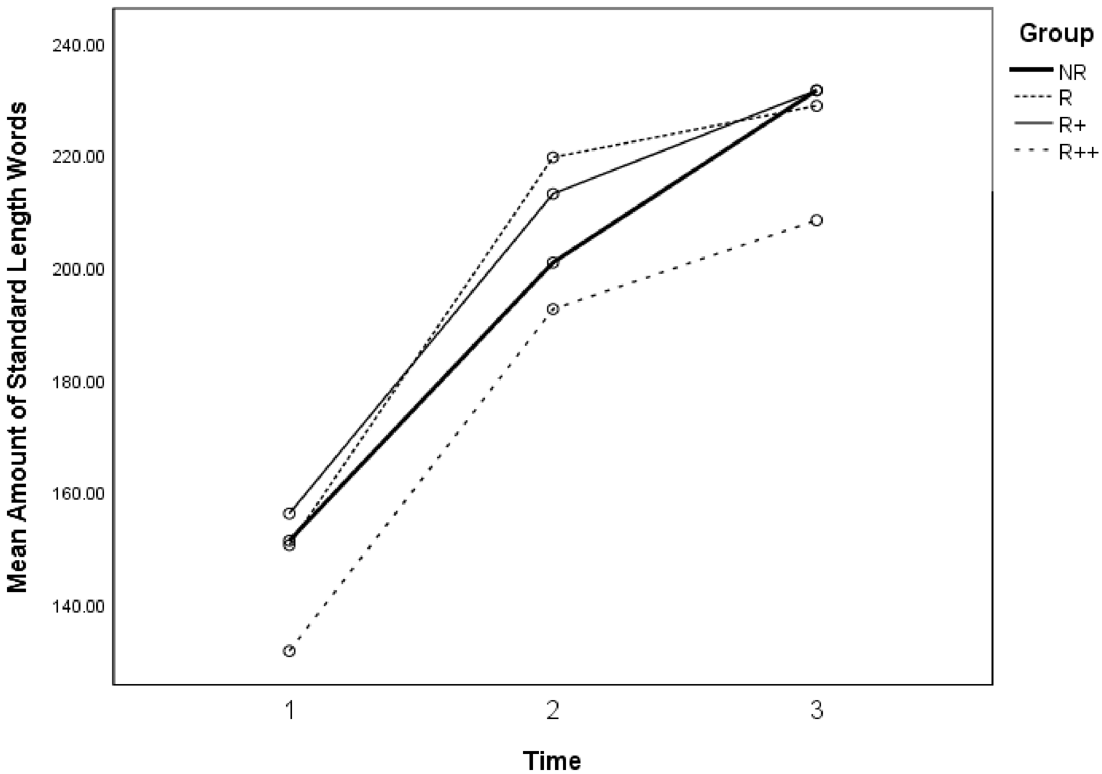

Figure 22. Mean number of standard length words for the groups. NR = No Reminder group, R = Reminder group, R+ = Remind + notice group, R++ = Remind + use group.

Seven outliers were identified and removed from the analysis, as assessed by examination of studentized residuals for having values greater than ±3. The measures of lexical complexity were normally distributed, as assessed by Normal Q-Q Plots. The homogeneity of variances assumption ($p > .05$) and homogeneity of covariances assumption ($p = .44$) were met as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. Mauchly's test of sphericity indicated that the assumption of sphericity was met for the two-way interaction, $\chi^2(2) = 2.54$, $p = .28$. There was no statistically significant interaction between the intervention and time on amount of standard length words, $F(6, 334) = 0.79$, $p = .58$, partial $\eta^2 = .014$. Therefore, analyses of the main effects for time and group were performed.

The main effect of time showed a statistically significant difference in amount of standard length words at the different time points, $F(2, 334) = 154.77$, $p < .001$, partial $\eta^2 = .48$. Follow-up pairwise comparisons were run where reported 95% confidence intervals and *p*-values were Bonferroni-adjusted. The unweighted marginal means of amount of standard length words was 147.42 (SE = 4.43) at Time 1, 206.50 (SE = 5.25) at Time 2, and 225.01 (SE = 5.47) at Time 3. The mean amount of standard length words increased from Time 1 to Time 2 by 59.07, 95% CI [48.64, 69.51], $p < .001$, by 18.52, 95% CI [6.97, 30.06], $p = .01$, from Time 2 to Time 3, and by 77.59, 95% CI [66.18, 88.99], $p < .001$, from Time 1 to Time 3.

The main effect of strength of reminder showed that there was a non-significant difference in mean amount of standard length words between treatment groups, $F(3, 167) = 1.26$, $p = .29$, partial $\eta^2 = .022$.

Hypothesis 5: Relationship Between Vocabulary, English Proficiency, and Writing Scores with Perceived Usefulness

The fifth hypothesis stated that improved vocabulary learning and retention, higher vocabulary, English proficiency, and writing scores will result in the participants reporting greater perceived usefulness of the vocabulary study as it relates to their writing. This hypothesis was explored using data obtained from the interviews.

In order to explore this hypothesis, it was important to determine what participants believed contributed to writing quality. In the second interview, the 20 interviewees were asked what they thought they needed in order to write a good essay. Their responses are summarized in Figure 23.

After carrying out adequate research about the topic, using appropriate vocabulary was considered the second most important element of a good essay. As such, the interviews sought to illuminate the role the reminders played in the use of the target vocabulary in the writing tasks. Table 43 shows a summary of the participants' responses.

Figure 23. Interviewee responses (*n* = 20) regarding elements of a good essay.

Table 43. *Summary of Participant Responses Indicating Effect of Reminders on Overall Writing Quality (n =12)*

	(R)	(R+)	(R++)
Yes (moderate to large effect)	1	3	4
Yes (little effect)	1	1	—
No effect	2	—	—

Note. R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group.

The results shown in Table 43 appear to mirror the treatments where stronger reminders with increased vocabulary practice results in greater perceived effects on writing quality. All four interviewees from the R++ group responded that they felt that the reminders had a moderate to large effect in their use of the target vocabulary in the writing tasks. The R+ group largely responded similarly. Although two interviewees from the R group responded that there was no effect, most of the interviewees reported that they felt that the reminders had a moderate to large effect on their writing quality as operationalized by vocabulary use. In response to the probe about the reminders, the interviewees responded as follows. Transliteration of the Japanese used by participants and my English translation of these responses appear in brackets immediately after the original language of the participants.

01-R++: Before tell use NGSL words, I don't use NGSL words. So, after tell, I think that I have to use NGSL words …. In this box, write word. If this box, if no this box, I remember use word, difficult. But, there is this box, use NGSL words, how many NGSL words, I can see. (line 76)

02-R++: If this paper look for me, I write NGSL, pick up and write. Final draft 書いて、それを潰しながら書いて行く事を心がけました。 [kaite, sore o tsubushinagara kaite iku koto o kokorogakemashita: I wrote the final draft and I was careful to check off the words while I kept writing.] Maybe I don't do this. (line 123)

In her response, participant 02-R++ indicated that the reminders might have influenced her to use the vocabulary list more actively in her writing. She intimated that she might not have done this if the teacher had not reminded them to use the vocabulary. Similarly, participants 03-R, 05-R+, and 07-R+ recognized that the vocabulary was useful and that the reminders played a role in learning them.

03-R: なるほど [naruhodo: I see]. たしかに [tashikani: certainly], I felt these words effect in writing essay. It is easy to write essay using these words. I remember to use the words. (line 151)

05-R+: I use the words in my homework so I could remember the words easily because I want to use more. (line 167)

07-R+: If I have this vocabulary list, I don't have to research on dictionary so it is useful. (line 161)

The participants who responded that the reminders had no effect belonged to the R group. In their interviews, the participants indicated either they did not understand why the teacher was telling them to use the vocabulary or they felt that the reminder was superfluous because they already knew most of the words.

06-R: Teacher tell students …これを使って言うけど、なんでなのかなって [kore o tsukatte iukedo, nandenanokanatte: he said use this, I didn't understand why we had to do it. I don't know] (line 147)

13-R: This vocabulary are almost I know. So, sometimes I didn't know words.気にしてなかった [kinishitenakatta: I didn't care about the vocabulary]. (line 160)

For participants 06-R and 13-R, it is possible that the reminder, without integrated vocabulary-focused tasks, such as those included in the R+ and R++ treatments, made the exercise less meaningful and, thus, resulted in some of the students failing to follow the instructions.

To sum up, the reminders had an effect on the participants' overall writing quality. The participants in groups with the stronger reminder treatments, the R+ and R++ groups, responded more positively. Their responses might have been due in part to the explicit vocabulary-focused tasks that were part of their essay-revision process. Having only a reminder to use vocabulary (R group) resulted in some students not fully understanding its purpose.

The interviewees were asked in both interviews to ascertain as to whether the reminders, and subsequent lack thereof, led to a change of perceived usefulness of the vocabulary study as it relates to their writing assignments in content course 1 and 2 (CC1 and CC2, respectively). Following Hatch (2002), the student responses were analyzed to identify patterns, themes, relationships, and explanations. Table 44 is a summary of their responses in both interviews and notes if there was a difference in their responses from Interview 1 to Interview 2.

Table 44. *Interviewee Responses Regarding Usefulness of the Vocabulary Study as it Relates to the Writing Assignments*

Participant number	Interview 1	Interview 2	Change
01-R++	Moderate	Useful	Positive
02-R++	Useful	Useful	
11-R++	Useful	Useful	
14-R++	Moderate	Useful	Positive
05-R+	A little	Useful	Positive
07-R+	Not useful	Not useful	
09-R+	Moderate	Moderate	
10-R+	Moderate	Useful	Positive
03-R	A little	Useful	Positive
06-R	A little	Not useful	Negative
13-R	Useful	Useful	
19-R	Useful	Moderate	Negative
04-NR	Useful	Useful	
08-NR	Not useful	Moderate	Positive
12-NR	Moderate	Useful	Positive
15-NR	Not useful	Moderate	Positive
16-NR	Moderate	Moderate	
17-NR	A little	Useful	Positive
18-NR	Useful	Useful	
20-NR	Moderate	Useful	Positive

Note. The interviewees are identified by an individual code and then by their group, 01 = Participant 01, R++ = Remind + use group, R+ = Remind + notice group, R = Reminder group, NR = No Reminder group.

Interviewees 03-R, 05-R+, 08-NR, 15-NR, and 17-NR showed the biggest changes in reported usefulness of the vocabulary study. Their reasons tended to be non-specific, but they reflected a belief in the general usefulness of the vocabulary.

03-R: I wasn't careful VC2 when I was studying vocabulary because CC2 teacher didn't say these words. But, these words can use everywhere. Writing report, speaking English with teacher so I useful these words. (line 209)

05-R+: でも使いやすいとは思うし、役立ったと思います [demo tsukaiyasui to ha omoushi, yakudatta to omoimasu: but I think it was easy to use and I think it helped]. And useful. 結構使ってたかも [kekkou tsukatteta kamo: I think I used a lot]. (line 127)

Interviewer: Was the NGSL useful for you?

08-NR: Maybe. (line 124-125)

15-NR: I guess they were useful but I'm not sure. (line 138)

Participant 17-NR elaborated further by explaining that she found it difficult to remember difficult words. These words are very likely low frequency, academic vocabulary that was encountered in Content Course 2 (CC2) through in-class lectures or assigned readings. In comparison, she found it easier to remember easy words implying that the NGSL was more useful.

17-NR: CC2 recount and narrative and report, I must remember all and rewrite. So, difficult word was I can't remember. So, many easy word is more remember. (line 82)

Overall, it can be seen that the participants who said that the vocabulary study was useful in Interview 1 tended to respond in the same way in Interview 2. Almost all other participants reported an increased in the perceived usefulness of the vocabulary study as it relates to their writing assignments in Interview 2. Only two participants, 06-R and 19-R, both from the R group, were considered to have a negative change from the first to the second interview.

Participant 06-R's change in opinion was from *a little* to *not useful* despite expressing a desire to learn. During the interviews, she revealed that she has already developed her own study habits and was trying to get good scores on different standardized English tests. The reason she felt that the vocabulary study in her English classes was not particularly useful was because the specific vocabulary presented in the class was not aligned with her own personal language goals.

06-R: TED app [a smartphone app] for listening. And word appli, TOEIC TOEFL appli 600 words. 学校の行き帰りの電車で [gakkou no ikikaeri no densha de: on the train on the way or going back from school] 高校の時から、高校がけっこう勉強のコースで、もうついていくので必死で、高校の時も通学の時はリスニングのテープを聞いていた。それがもう習慣 [koukou no tokikara, koukou ga kekkou benkyou no ko-su de, mou tsuiteiku no de hisshide, koukou no toki mou tsuugaku no toki ha risuningu no te-pu wo kite ita. Sore ga mou shuukann: My high school course was quite high level so I had to study a lot just to pass. I usually listened to the English listening tapes. It has been my habit (since high school)]

In the first interview, participant 19-R reported that the vocabulary study helped her a lot with the writing assignments and that she tried to use the target vocabulary in every writing. However, in the second interview, she reported that the vocabulary study was of moderate usefulness (a 2 out of 5 rating for usefulness). The reason for this change in opinion can be attributed to participant 19-R already having established a preferred method of vocabulary study. She reported that she uses a *tangocho* (vocabulary book) for studying vocabulary. These books are commercially available publications containing vocabulary lists for a predetermined purpose such as test-preparation. Interestingly, despite reporting a preference for the *tangocho*, she also stated that she doesn't like using it because *it's like study* (line 45) and generally does not like studying vocabulary because *it's just memorize* (line 51).

Interviewer: How do you study vocabulary? What is your vocabulary study style?

19-R: 単語帳 [tangocho: vocabulary book]

Interviewer: Okay. All right. So this is your study style for vocabulary.

19-R: Yes. Or, as I said, website.

Interviewer: So, why do you use a 単語帳 [tangocho: vocabulary book] to study vocabulary?

19-R: 単語帳 [tangocho: vocabulary book] is useful because English and Japanese and example. So, just useful.

Interviewer: Sure. Did you use the Quizlet or the Moodle quizzes from VC?

19-R: Yes, yes.

Interviewer: Okay. But you still prefer the 単語帳 [tangocho: vocabulary book]? I'm wondering how much you use for the different things.

19-R: Quizlet and Moodle, I used very short.

Interviewer: Short time?

19-R: Yeah. Because it was interesting but I can't continue.

Interviewer: Okay. So, 単語帳 [tangocho: vocabulary book] sounds like the one you use the most.

19-R: Yes. (lines 58-78)

One participant, 07-R+, reported in both interviews that she felt the vocabulary study was not particularly useful. In her case, she said that she didn't find the vocabulary study useful because she already knew the words. Therefore, the additional reminders and vocabulary focus was inconsequential to her.

07-R+: Because almost I know these words so … not so much.

Interviewer: How frequently did you try to use words from here [pointing to the NGSL list] in your writing?

07-R+: From column 1 to 4.

Interviewer: Column 1 to 4?

07-R+: So, I remember to use, I have to use those words, and from column 5, I didn't mind or care about those.

Interviewer: Okay. Why not?

07-R+: Because it didn't match what I wanted to use.

Interviewer: It didn't match the topic?

07-R+: Yeah. Writing report or narrative, I didn't use those words because it didn't match. (lines 199-210)

This sentiment was repeated in the second interview.

07-R+: Because some words matching to our topic. Writing recount or narrative but many of them didn't match. And I can use more easy words like similar meaning so I don't have to use those words. So, I didn't use. (line 120)

And later in the same interview, she said:

07-R+: Sometimes it was useful but the topic was different, those didn't match my essay. I can't use. If I must use those vocabulary, I have to use so I use. But 強制されたら書く [kyousei saretara kaku: If I was forced, I would use]. (line 208)

In general, the interviewees reported an increased perceived usefulness of the vocabulary study as it relates to their writing. In the two cases where there was a decrease in perceived usefulness, it might be explained by the participants already having an established and preferred study style. For the one participant who did not find the vocabulary study useful in both interviews, she attributed this to already knowing the target vocabulary.

CHAPTER 6

DISCUSSION

In this chapter, I restate the hypotheses and briefly summarize the results presented in Chapter 5, before interpreting the findings in the light of existing research. Following this, I discuss the pedagogical implications with regards to the effect of reminders on transfer of learning.

Hypothesis 1: Salience of the Reminders and Effect on Study Time

Hypothesis 1 stated that the reminders would be more salient in the stronger treatments due to the inclusion of tasks designed to facilitate noticing the target vocabulary and guidance towards their use in writing. The three reminder conditions included tasks with increasing processing depths, reflected by progressively higher Technique Feature Analysis scores (Nation & Webb, 2011) (see Appendix V). These tasks were designed to promote the noticing of gaps in the participants' linguistic ability. As a result, it was hypothesized that the reminders would be more salient for participants in stronger treatment conditions than those in weaker treatment conditions. In turn, the participants in stronger treatment conditions would spend more time studying the target vocabulary. The reminders were considered salient if the interviewees mentioned them while describing a typical content course lesson. The amount of study time was assessed by examining the Quizlet and Moodle data for each participant, and their self-study logs. The study time data were analyzed using the Jonckheere-Terpstra trend test to evaluate if there were significant differences between the four conditions.

Summary of the Results: Salience of the Reminders and Effect on Study Time

Salience of the reminders was evaluated using the interview data. The 20 interviewees were asked to describe a typical content course lesson. Of the 12 participants who were in the treatment groups, only six specifically mentioned that their teacher reminded them to use the target vocabulary. Of those six, two were from the R++ group, three were from the R+ group, and one was from the R group. While this might indicate that some of the participants did not notice the reminders, it should be considered that the lack of mentioning the reminders could be a result of a failure on my part to adequately probe during the interviews. That being said, the reminders appear to have been salient. The participants in the three treatment groups indicated some degree of conscious and deliberate use of vocabulary learned in one class in a second, separate class and the finding confirms the hypothesis that the reminders would be most salient to students given the strongest reminders.

The results of the Jonckheere-Terpstra trend test showed that the amount of time spent on vocabulary study was significantly affected by the degree of reminders and that there were significant differences between the R++, R and NR groups for the self-study logs in semester 1 (SSLog 1). However, the Jonckheere-Terpstra trend test reported a trend of descending medians, meaning that the amount of study time fell as the treatment strengthened, contrary to the hypothesis. Furthermore, the results from analyzing the self-study logs in semester 2 returned non-significant results.

Analysis of the Quizlet (2017) and Moodle (2016) data showed non-significant results indicating that the reminders did not affect the amount of time spent studying using these two resources. It must be noted that use of Quizlet was optional and that many participants elected not to use it as doing so did not directly affect their grades. This resulted in large gaps in the data. Moodle use, however, was part of the Vocabulary Course's grading criteria and participants were told that completing the weekly practice quizzes would account for 10% of their final grade. Compared to the Quizlet data, the Moodle datum was more complete but there were still gaps as some participants might have considered 10% inconsequential to their final grade.

Interpretation: Salience of the Reminders and Effect on Study Time

Evaluating the salience of the reminders was important because a failure to notice or understand the treatment would mean that any observed changes in study time, vocabulary learning, vocabulary retention, or test scores could not be attributed to the treatment. The interviews revealed that the reminders were generally salient to the participants, particularly those in stronger treatment conditions. This result suggests that administering reminders can be an important first step towards helping learners notice gaps in their linguistic ability.

It was then hypothesized that the participants in stronger treatment conditions would spend more time studying the target vocabulary than those in weaker conditions. Increased study time would allow the participants to benefit from spacing effects (Baddeley, 1997; Bahrick & Phelps, 1987; Mizuno, 2003; Nakata, 2015) and potentially result in improved vocabulary learning and retention. Previous research in learning and memory has shown evidence that spaced learning is a more efficient learning strategy than crammed learning (Dempster, 1987; Kornell, 2009). However, many learners employ cramming, particularly just before a test. Given that the participants were tested on the target vocabulary three times during semester 1, it was predicted that the participants would report substantially more time on vocabulary study in the week or day just prior to each test. In the current study, study time was measured in three ways: (a) the amount of Quizlet (2017) use, represented as the number of times a study mode was started or completed, (b) the total number of practice quizzes taken on Moodle (2016), and (c) a weekly study log in which the participants reported the amount of time they spent studying vocabulary using Quizlet or other means throughout the previous week and the previous day. These measures were analyzed using the Jonckheere-Terpstra trend test to evaluate if study time varied across the groups in this study. The Jonckheere-Terpstra test revealed that the participants in stronger treatment conditions spent less time studying the vocabulary than those in weaker conditions, contrary to the hypothesis.

Overall, care must be taken when interpreting the finding that study time decreased in stronger treatment conditions. Self-reported measures such as the self-study logs are known to be flawed sources of data as they depend solely on participants' memory. Herrmann (1982) noted that using memory questionnaires should as the sole method of assessment is inadvisable as the strength of the validity data is too low. Even for relatively recent events, memory is not perfect and discrepancies should be expected. To illustrate this point, Zuriff (2003) found that participants overestimated the amount of time spent studying by more than 40% in self-reported time estimates compared to the data recorded in their time diaries. Another factor that possibly affected the reliability of the self-report in this study was participant honesty. Meleis and Dagenais (1980) described that self-report data have inherent risks as respondents can create false impressions, consciously or unconsciously. Furthermore, they found that positive attitude played a role in this form of bias. Therefore, despite being told their responses would not affect their grades, it is possible that the participants in this study distorted the data due to a desire to create a good impression or through an overestimation of their study time.

The inclusion of data regarding Quizlet (2017) and Moodle (2016) data were an attempt at mitigating the problems associated with self-reports. However, the zero-stakes nature of Quizlet and low-stakes penalty for not using Moodle led to large gaps in the data, which is likely to have contributed to the non-significant findings. Further compounding the problem was that the use of Quizlet and Moodle data were limited to frequency counts rather than duration of use. This had the effect of giving more weight to participants who might have had many short Quizlet/Moodle sessions over those with fewer longer sessions. Therefore, the results regarding effect of reminders on study time are inconclusive.

Hypothesis 2: Effect of Study Time on Vocabulary Learning and Retention

Hypothesis 2 stated that the participants in stronger treatment conditions would show improved vocabulary learning and retention than those in weaker conditions because they would have spent more time studying the target vocabulary. The NGSLT-Receptive and NGSLT-Productive were administered at three times in order to obtain evidence of vocabulary learning and retention. Hypothesis 2 was explored by conducting correlation analyses between the participants' study time and NGSLT-Receptive and NGSLT-Productive Rasch logits to evaluate if there was a relationship between study time, vocabulary learning, and vocabulary retention. Correlation analyses were also conducted between the participants' study time and amount of target vocabulary used in Tracking Essay 2. Only Tracking Essay 2 was used in the analysis as Tracking Essay 1 was written before the treatments were administered and Tracking Essay 3 was written one semester after the treatments had ceased. Hypothesis 2 was further explored by identifying the target vocabulary that the participants used in the tracking essays, and through examination of the participants' Vocabulary Inventory Surveys.

Summary of the Results: Effect of Study Time on Vocabulary Learning and Retention

The results from analyzing the participants' NGSLT-Receptive and NGSLT-Productive Rasch logits, target vocabulary use, and study time did not support hypothesis 2. A mixed ANOVA showed that there were significant differences in NGSLT-Receptive and NGSLT-Productive Rasch logits from Time 1 to Time 2, and from Time 2 to Time 3. This result indicated that all the participants performed better on the receptive and productive vocabulary tests throughout the data collection period. However, the mixed

ANOVA also showed that there were no significant differences between the treatment conditions indicating that the strength of reminders did not affect the rate of vocabulary learning or retention.

Identifying the target vocabulary used in the tracking essays returned similar results. One-way ANOVAs for each of the four conditions revealed that there were non-significant effects of reminders on vocabulary learning or retention. However, despite the non-significant results, the mean number of identified target vocabulary that was categorized as learned or retained generally paralleled the strength of reminders, suggesting a possible effect.

Correlation analyses between study time and the vocabulary tests showed that Moodle use significantly and positively correlated with the NGSLT-Receptive ($r = 0.28$) and NGSLT-Productive ($r = 0.21$) at Time 2, and with the self-study logs ($r = 0.30$) at Time 3. The correlation coefficients indicate that the measures are not closely related. Analysis of the Vocabulary Inventory Surveys showed that there were no significant differences between the groups regarding the amount of known vocabulary from Time 1 to Time 2. Coupled with the finding that many of the correlations were insignificant, or small when they were significant, this suggests that the reminders had a negligible effect on vocabulary learning.

Analysis of the amount of target vocabulary use in Tracking Essay 2 did not show significant correlations with any of the self-study measures. However, a follow-up correlational analysis between the amount of target vocabulary use and cumulative vocabulary test scores showed a small but significant result (r = .220, p = .003). While the strength of association was small, the result showed that there is a relationship between vocabulary test scores and vocabulary use in writing, which is unsurprising as a higher test score shows evidence of a larger lexicon and, thus, potentially greater use of the target vocabulary than those with a smaller lexicon. A possible explanation for the small strength of association is that the vocabulary tests assessed the whole range of target vocabulary whereas the Tracking Essay placed no emphasis on using particular words. Instead, participants were free to use any words that they felt were appropriate. Furthermore, there were no penalties for not using the target vocabulary.

Interpretation: Effect of Study Time on Vocabulary Learning and Retention

Overall, the results showed that vocabulary learning was occurring in each of the four conditions. This is attributed to the curriculum at the institution and was expected. However, the results also indicated that the reminders did not have an effect on vocabulary learning and retention as evaluated by the NGSLT-Receptive, NGSLT-Productive, target vocabulary used in the tracking essays, and Vocabulary Inventory

Surveys. This finding is surprising as the results from hypothesis 1 suggested that the reminders were generally salient but resulted in less study time for stronger conditions.

Considering the results together, it appears that the participants in stronger treatment conditions were able to learn and retain vocabulary despite reporting less study time. This finding suggests that the participants might have learned the vocabulary elsewhere and, as such, there was less need for independent study time. Indeed, if we consider that the strongest treatment group had to identify useful vocabulary and write them above their essays, this likely resulted in the participants learning these words during the course of the treatment, that is, during their lessons and not necessarily during extracurricular study time. In the NR condition, the participants would only be exposed to the target words in the vocabulary course, while revising or preparing for the tests. However, the participants in the reminder treatments would have the opportunity for more frequent exposure to the words due to the tasks. The R+ treatment participants had to refer to a truncated vocabulary list and select potentially useful vocabulary while revising their essay. The R++ treatment participants were given a more complete list, select potentially useful vocabulary, and write those words in a provided space above the final draft of their essay. The greater opportunities for exposure to the target vocabulary and the increased involvement required of the tasks would allow the participants to benefit from frequency effects (Bybee, 2009; N. Ellis, 2015; Wulff & Ellis, 2008). As a result, having learned those words, they would not have had to devote time to study of them in preparation for the vocabulary tests. This essentially confirms prior research that show how greater involvement loads (Hulstijn & Laufer, 2001) and experience through use (Tomasello, 2000) can improve long-term retention. This is an important finding because, if curriculum designers consider how various courses in a coordinated program can be connected, they might find it reinforces overall learning. For students, it is potentially motivating to understand how a task connects to their overall language development and has the added benefit of reducing study time.

Research regarding long-term memory and learning (e.g., Laufer & Hulstijn, 2001; Nation & Webb, 2011) has shown that, provided certain conditions are met (i.e., fulfilling learning needs, noticing new information, and considering appropriate options), we can expect language learners to successfully use the vocabulary that they are studying. In this study, the stronger treatments (R+, R++) required the participants to identify words that were potentially useful for their homework and, in the case of the strongest treatment, write them in a space provided on their assignment sheets. It was hypothesized that this attention to the vocabulary would lead participants in these groups (R+, R++) to show better learning, retention, and correct use of the target vocabulary than those in the comparison (NR) and weaker treatment (R) groups. However, the results indicated no significant differences. This finding suggests that the tasks were insufficient in fulfilling the conditions required which contributed to the non-significant results.

It is also important to consider the role of topic in a writing task. Analysis of the target vocabulary used also revealed that there was a larger number of unused vocabulary (previously identified as learned, but not used again in a later essay) as the treatment strength increased. A follow-up examination of the tracking essays revealed that, out of 178 participants, 32 (17.98%) wrote about the same topic in all three tracking essays, 65 (36.52%) wrote about the same topic in two tracking essays, and 81 (45.51%) wrote about different topics in each of the three tracking essays. This finding is important because it has been established that the type of vocabulary that should be used is dependent on the topic of the writing task (see Chapter 2: Components of Good Writing). Given that only 18% of the participants wrote about the same topic in all three tracking essays, it can be concluded that the inconsistency in writing topic by the remaining 82% added an extra variable in the analysis that could not be controlled. The variability in writing topic likely distorted the estimates of learned and retained vocabulary obtained from analyzing the tracking essays.

Overall, most of the self-study measures and vocabulary tests did not show significant correlations with each other and none of the correlations were large. At both Time 2 and Time 3, Moodle (2016) use correlated significantly with the observed vocabulary test scores but at both times, the correlation was small. The association was expected as Moodle use was a part of the course grade and both the Moodle quiz and NGSL tests are vocabulary tests. However, the small strength of association was unexpected. Furthermore, despite Quizlet (2017) also being a vocabulary test, its use was optional, and many participants elected not to use it. The self-study logs correlated significantly with Moodle use only at Time 3 but not at Time 2. One possible explanation for this inconsistency is that the self-study logs were based on the participants' ability to recollect the amount of time they spent engaged in vocabulary study throughout the previous week which would have made the data unreliable (Herrmann, 1982; Zuriff, 2003). Furthermore, it is possible that some of the participants, either consciously or unconsciously, wanted to create a good impression by overestimating their study time (Meleis & Dagenais, 1980). The same issue was observed when examining the results of the correlation analyses between study time and target vocabulary use. None of the comparisons were significant indicating that study time was not associated with the amount of target vocabulary use in a writing task.

The small but significant result of the follow-up correlation analysis between the amount of target vocabulary use and cumulative vocabulary test scores showed a positive relationship between vocabulary test scores and vocabulary use in writing; a higher test score shows evidence of a larger lexicon and, thus, potentially greater use of the target vocabulary than those with a smaller lexicon. A possible explanation for the small strength of association is that the vocabulary tests assessed the whole range of target vocabulary whereas the Tracking Essay placed no emphasis on using particular words. Instead, participants were free to use any words that they felt were appropriate. Furthermore, there were no penalties for not using the target vocabulary.

Hypothesis 3: Effect of Vocabulary Learning and Retention on GTEC and Writing Scores

Hypothesis 3 stated that improved vocabulary learning and retention in stronger conditions would result in better GTEC and tracking essay scores. This hypothesis was based on studies that have shown that L2 vocabulary size is a good predictor of L2 reading comprehension (Coady, Magoto, Hubbard, Graney, & Mokhtari, 1993; Koda, 1989; Laufer, 1997). Therefore, L2 readers need to have immediate and automatic access to at least 3,000 word families for good reading comprehension (Laufer, 1997). The target vocabulary range of the first 1,000 most frequent word families as listed by the NGSL (Browne, Culligan, & Phillips, 2013b) is well within the 3,000 word family requirement, which means that the findings of this study would be unaffected by ceiling effects.

It was also hypothesized that improved vocabulary learning and retention would expediate the transition from receptive to productive vocabulary. Studies have shown that vocabulary is a major predictor for writing scores (e.g., Astika, 1993; Jordan, 1997; Santos, 1988). Therefore, the ability to use vocabulary more freely would result in improved writing scores. Hypothesis 3 was explored by conducting correlation analyses between the participants' NGSLT-Receptive, NGSLT-Productive, GTEC scores, and the FACETS fair averages of their tracking essays.

Summary of the Results: Effect of Vocabulary Learning and Retention on GTEC and Writing Scores

The correlation analyses showed that the GTEC, NGSLT-Receptive, NGSLT-Productive, and FACETS fair averages of the tracking essays had significant, positive correlations with medium to large strengths of association. This result was expected as the four measures assess language proficiency either in part or in whole. At Time 1, the GTEC was more highly correlated with the NGSLT-Receptive ($r = .64$) than the NGSLT-Productive ($r = .53$). At Time 3, the NGSLT-Receptive and NGSLT-Productive had equal strengths of association with the GTEC ($r = .59$). Analysis of the GTEC scores using a mixed ANOVA returned significant results that the participants in all conditions improved from Time 1 to Time 3. However, there was a non-significant difference in mean GTEC scores between the conditions, suggesting that the strength of reminders did not affect the rate of improvement on the GTEC. At each time, the GTEC had a medium strength of association with the NGSLT-Receptive and NGSLT-Productive.

Interpretation: Effect of Vocabulary Learning and Retention on GTEC and Writing Scores

Hypothesis 3 was concerned with the potential effect different reminder treatments had on depth of vocabulary knowledge and, by extension, on overall language proficiency. The correlation analyses of the NGSLT-Receptive, NGSLT-Productive, GTEC, and tracking essay scores showed that, in general, participants who scored highly on one measure tended to score highly on others. This is unsurprising and is in line with previous research showing the relationship between L2 vocabulary with reading comprehension (Coady, Magoto, Hubbard, Graney, & Mokhtari, 1993; Koda, 1989; Laufer, 1991) and writing scores (e.g., Astika, 1993; Jordan, 1997; Santos, 1988). The analyses showed that the NGSLT-Receptive and NGLST-P had approximately equal strengths of association ($r \geq .60$) indicating that the participants were performing similarly on each administration of the tests. In addition, both the NGSLT-Receptive and NGSLT-Productive have roughly the same predictive power for writing scores. Analysis of the GTEC scores showed that the gains were statistically significant. Results of a mixed ANOVA showed that the strength of reminders did not have an effect on the rate of improvement on the GTEC. However, it is interesting to note that the NGSLT-Productive and NGSLT-Receptive had equal strengths of association with the GTEC at Time 3 ($r = .59$). At time 1, the NGSLT-Productive ($r = .53$) had a lower strength of association with the GTEC than the NGSLT-Receptive ($r = .64$). These results show that the participants' productive vocabulary knowledge is sharing more variance at Time 3 (34.81%) than at Time 1 (28.09%). This finding suggests that the participants' depth of vocabulary knowledge improved between Time 1 and Time 3.

Overall, the results for hypothesis 3 showed a positive correlation between vocabulary learning and retention with GTEC and writing scores. Simultaneously, the results for hypothesis 2 indicated that vocabulary learning was occurring, despite a non-significant result for the effect of reminders. These results, together, suggest that improved vocabulary learning has a positive effect on GTEC and writing scores. The lack of significant difference between the four groups suggests that reminders are just one of many factors that have an impact on the GTEC and writing scores.

Hypothesis 4: Effect of Vocabulary Learning and Retention on Accurate Target Vocabulary Use in Writing

Hypothesis 4 stated that improved vocabulary learning and retention for the reminder groups would enable the participants to use the target vocabulary more accurately in the tracking essays. As with hypothesis 3, hypothesis 4 was based on previous studies that showed that L2 vocabulary size is one of the main predictors of writing scores (Astika, 1993; Jordan, 1997; Santos, 1988). However, vocabulary has to be used accurately and effectively. The identified target vocabulary were evaluated using a chi-square test of independence to examine the differences in amounts of correct usage in the four conditions.

This hypothesis was further explored by evaluating the participants' tracking essays using the JACET writing rubric to evaluate writing quality. In addition, the lexical complexity and the number of standard length words in each Tracking Essay were calculated. As the definition of *good* writing varies from person to person, these three instruments were selected in an effort to quantify writing quality more completely. These data were analyzed using a mixed ANOVA to investigate the effect of strength of reminder on correct use of the target vocabulary in a writing task for the four conditions.

Summary of the Results: Effect of Vocabulary Learning and Retention on Accurate Target Vocabulary Use in Writing

Regarding correctness of use, the participants' tracking essays were analyzed using a chi-square test for independence, which returned mixed findings (see Table 38). The target vocabulary that was identified as *learned* were evaluated as being used correctly, partially correctly, or incorrectly in terms of grammar and contextual fit. The chi-square analysis showed that stronger reminder treatments led to more instances of correct usage and also resulted in the treatment groups (R, R+, and R++) having higher counts for correct usage than expected by the chi-square procedure. In contrast, the comparison (NR) group had a lower count. For partially correct use of the target vocabulary, the comparison (NR) group had more instances of this whereas the treatment groups had fewer. For incorrect use of the target vocabulary, there were no clear trends in the data between the comparison and strongest treatment groups. The results also showed that there were no significant differences between the conditions for amount of target vocabulary used correctly, partially correctly, or incorrectly.

For the participants' JACET writing rubric scores, expressed as FACETS fair averages, the mixed ANOVA revealed that, while all conditions showed significant increases in fair averages over time, there were no significant differences between conditions. Similar results were found from the analyses of the lexical complexity and number of standard length words of the tracking essays. Overall, analysis of the tracking essays showed that the participants in all conditions demonstrated improved writing quality as measured by the JACET writing rubric, that they were writing more lexically complex essays, and that they were writing more words in the same time period.

However, the reminders did not significantly affect the rates of improvement for any of the three measures of writing quality.

Despite the non-significant results of reminders on writing quality, there was an indication that the rate of improvement for lexical complexity and standard length words were different across the groups. Figures 20 and 21 in Chapter 5 shows the progression of mean estimates of lexical complexity and standard length words. In both figures, the gradients for the comparison (NR) group is approximately constant from Time 1 to Time 2 to Time 3. However, the gradients for the treatment groups all show a steeper slope from Time 1 to Time 2, followed by a shallower slope from Time 2 to Time 3.

Interpretation: Effect of Vocabulary Learning and Retention on Accurate Target Vocabulary Use in Writing

Hypothesis 4 was concerned with the effect of the reminders on accurate use of the target vocabulary in a writing task. The results from the chi-square analysis indicated that the participants in the treatment conditions were more likely to use the target vocabulary correctly with fewer instances of partially correct or incorrect usage. Furthermore, the contingency table (see Table 38) showed that the participants in the three treatment conditions used more of the target vocabulary than expected by the chi-square analysis. These findings lead to the conclusion that hypothesis 4 was confirmed.

A wide variety of materials and tasks are used in language classrooms with the overall aim of language development. Second language educators teach with the hope that lexis will be used in other situations. In this study, treatments were administered in one course to investigate if vocabulary learning could be influenced so that it would have an impact on the use of vocabulary in writing tasks in a different course.

The results of the mixed ANOVA help explain the degree to which the reminders, through their effects on vocabulary learning, contributed to overall writing quality. Writing quality was considered an important measure to include in the overall analysis as it provided the clearest manner in which to observe the impact of the treatments on a secondary variable. In evaluating student essays, teachers typically use a rubric to determine if the relevant criteria have been met. Rubrics vary from simple, short checklists, to complex forms that require careful examination of each essay. However, rubrics adopted by teachers might not include lexical complexity or standard length words as criteria. On the whole, all the participants improved in their writing scores, lexical complexity estimates, and number of standard length words produced. The finding that the participants in the NR condition also showed significant gains suggests that instruction and practice led to improvement. This result is not surprising as it is reasonable to expect improvements in writing quality, lexical complexity, and perhaps even writing fluency from regular instruction and practice.

The continued improvement by all groups suggests that, on the whole, the participants wrote more better essays through the year. However, the slight decrease in mean lexical complexity by the R group was unexpected. It is possible that these participants were using the target vocabulary while writing their class-assigned essays but did not employ this vocabulary under test conditions. This would have the effect of reducing their lexical complexity score. Similarly, if they tended to repeat many words or write simpler sentences, this would have had the same effect. Also, it must be considered that many of the participants did not choose to write about the same topic for each Tracking Essay which would have affected the vocabulary choice and use.

Overall, the results of the repeated-measures ANOVA for the three measures of writing quality provide insights into the secondary effects of reminders on learning. It seems that there are directly observable benefits that, given their relative ease, would make it a simple decision to include reminders in courses for a coordinated program. However, these results must be interpreted with some degree of caution as there are indications that other factors might be influencing them.

Hypothesis 5: Relationship Between Vocabulary, English Proficiency, and Writing Scores with Perceived Usefulness

Hypothesis 5 concerned how the participants perceived the usefulness of the reminders with regard to their writing. In essence, this question was primarily about whether the participants were aware of the connection between the vocabulary study in the vocabulary course and the writings in the content course. It was hypothesized that the reminders would be more salient to participants in stronger treatment conditions (hypothesis 1). As a result, those participants would be more cognizant of the potential benefits of the vocabulary study with regards to their content class assignments. In motivational research, the Hackman and Oldham model (1980) describes meaningfulness as a crucial component of intrinsic motivation. In short, a task that had a clear purpose would be more intrinsically motivating.

The interviewees were asked about this perceived usefulness in both interview sessions, once at Time 2, and again at Time 3. Throughout the two semester data collection period, the assignment to study vocabulary and instruction in various genres of writing were constant. As such, this question sought to determine if there was a change in perceived usefulness of the vocabulary study throughout that two-semester period.

Summary of the Results: Relationship Between Vocabulary, English Proficiency, and Writing Scores with Perceived Usefulness

The results of the analysis for the first interview showed that the interviewees generally perceived the vocabulary study to have been moderately useful to their writing. Six out of 20 interviewees (two each in the NR, R, and R++ groups) reported it as having been useful, with a further 11 reporting the activity as being moderately useful or a little useful. The remaining three interviewees (two in the NR group, one in the R+ group) commented that the vocabulary study was not useful. It is interesting to note that the participants in stronger treatments generally responded more favorably of the vocabulary study when compared to those in weaker treatments.

The data from the second interview generally revealed a greater perceived usefulness of the vocabulary study compared to the first interview. Ten participants, who previously reported that the vocabulary study was moderately useful, a little useful, or not useful, responded more favorably in Interview 2. Eight reported the same perceived usefulness in interviews 1 and 2, five of whom said "useful" both times. Only two interviewees' (06-R and 19-R) comments showed that they felt the vocabulary was less useful in Interview 2 than in Interview 1. Through probing in the interviews, these two participants implied that they felt the vocabulary study was not particularly useful for their writing as they were already using previously established study habits. One of the participants (06-R) commented further that she was focusing on studying for a variety of standardized tests and, as such, the vocabulary study that she was directed to do was not a good fit for that purpose. Only one interview participant (07-R+) reported in both interviews that the vocabulary study was not useful. Analysis of her transcripts showed that in both interviews, she felt that she did not find the vocabulary study useful as she already knew the words. As a result, the reminders and subsequent vocabulary-noticing task was inconsequential.

The interviews also provided insight into the effect of the reminders on perceived writing quality. The participants in the stronger treatment conditions reported more positive perception of their writing than those in the NR group. This is attributed to the vocabulary-focused tasks in the R+ and R++ groups as the tasks enabled the participants to notice deficiencies in their language, which then prompted them to address these areas (Swain, 1985). Moreover, when these issues were considered resolved, it likely led to a sense of accomplishment. The interview data also revealed that the participants viewed vocabulary as second only to research requirements for good essays. In short, participants who were asked to identify potentially useful vocabulary and/or to make a note of these words during the revision process reported that the reminders had a positive effect on their writing quality.

Interpretation: Relationship Between Vocabulary, English Proficiency, and Writing Scores with Perceived Usefulness

The fact that many students said that they felt vocabulary is useful was expected. Indeed, language educators and learners alike inherently understand the value of vocabulary. The participants reported in the interviews how vocabulary, aside from adequate research, was one of the most important factors in good writing. This is consistent with professors' comments about the seriousness of lexical errors in students' written work (Astika, 1993; Jordan, 1997; Santos, 1988). However, learners are used to tests being the typical outcome for vocabulary study. This results in a compartmentalization of learning in that the test becomes the primary driver for vocabulary study. As such, learners' study methods are likely to suit this purpose.

The interview data revealed two patterns. First, stronger treatments typically resulted in the interviewees reporting that the vocabulary study was useful for their writing assignments. Second, the reported usefulness of the vocabulary study changed over the course of two semesters, with perception becoming more positive across all groups. These two findings suggest that it takes time for students to appreciate the usefulness of tasks in a language classroom and is expedited through the use of attention-raising activities. In particular, when the study task (i.e., vocabulary study) and the purpose for the task (i.e., writing assignments) are in separate contexts, the usefulness of the study might not be readily apparent. This suggests it is incumbent on language instructors and curriculum designers to impress upon the learners the importance or relevance of study to their learning. While highly motivated students will always need little pushing, less motivated learners might benefit from additional reminders. However, it must be remembered that even highly motivated learners might choose to disregard such instruction particularly those with an established study style or where personal goals are not in alignment.

Pedagogical Implications

The results of this study offer one key pedagogical implication. Essentially, administering reminders to EFL learners has four main benefits. First, the results demonstrate that even simple reminders are salient. This finding is important because adding reminders to a lesson requires minimal effort but can have positive effects regarding overall language learning and motivation. Reminders can facilitate the noticing of linguistic gaps, which can serve as a guide for L2 learners to focus their attention on addressing these deficiencies (Swain, 1998). Coordinated language programs, especially, can exploit this finding by providing reminders at relevant times in the curriculum in order to clarify or bring attention to how various elements of the program are related. For EFL learners, many language skills are transferable across contexts. For example, vocabulary instruction in one course is expected to benefit the learners in other courses. So, reminding learners to use vocabulary in other courses can serve to illustrate potential uses of these words in a variety of situations. Similarly, learning how to write a report in an L2 writing course enables learners to write reports in non-EFL areas, such as in their academic discipline, provided the learners are aware of this transferability.

Second, pairing vocabulary-use reminders with tasks designed to encourage use of the target vocabulary resulted in less out-of-class vocabulary study time. The Technique Feature Analysis (Nation & Webb, 2011) framework suggests that depth of knowledge is positively affected by these tasks. The finding that the participants in stronger treatment conditions performed at least as well as those in weaker conditions, despite reporting less study time, indicates that the vocabulary learning was taking place during these in-class tasks. As a result, the learners in the stronger treatment conditions did not need to spend as much time outside class studying the vocabulary as those in weaker conditions. Therefore, incorporating tasks that are evaluated using Technique Feature Analysis can be beneficial in strengthening vocabulary knowledge while lessening the out-of-class time devoted to study by learners.

Third, this study also suggests that the reminders had an effect on the amount of target vocabulary used and degree of correct usage. The results indicated that the participants in the stronger reminder treatments used target vocabulary more frequently with more incidences of correct usage than those in weaker conditions. Therefore, instructors utilizing tasks that push learners to use the target language provides them with the opportunity to benefit from implicit learning (Bybee, 2009; N. Ellis, 2015: Wulff & Ellis, 2008).

Fourth, the reminders helped the participants in the stronger treatment conditions understand the value of the vocabulary study more quickly than those in weaker conditions. This finding suggests that, by showing how certain skills or tasks can be applied beyond the immediate context, learners will be able to understand the interplay of various linguistic functions.

Ultimately, the main takeaway is that reminders have the potential to positively affect learner motivation, which, in turn, has far-reaching ramifications regarding the learning of specific content in the classroom and beyond. For learners, the purpose of certain materials or tasks is not always clear. Demonstrating how to apply various skills or knowledge in different contexts can be beneficial to learners in that clarity of intent and purpose is provided, which can help them change their attitudes about learning. While it takes time and effort to develop a coordinated language program, educators and administrators can use reminders as an easy way to reinforce the elements of the program by illustrating how they are connected.

Summary

In this chapter I have discussed the results of the study, focusing on each of the hypotheses in turn, summarizing, and interpreting the results. In Chapter 7, I briefly summarize the main findings of the study, then discuss the limitations of the study, before making suggestions for future research examining transfer of learning.

CHAPTER 7

CONCLUSION

In this chapter I begin by briefly summarizing the main findings of the current study, before considering some of the limitations. Following this, I make several suggestions for future research before giving my final conclusions regarding the importance of vocabulary use reminders and the potential effect of various types of reminders on the transfer of knowledge.

Summary of the Findings

This study was conducted to investigate the effectiveness of vocabulary use reminders on L2 language learners' study time, vocabulary learning, language proficiency, and writing quality. This study produced four main findings. First, study time decreased as the strength of treatment increased. This was contrary to the hypothesis that stronger treatments would result in more vocabulary study time.

The second finding was that the participants in the treatment groups learned the target vocabulary at least as well as the comparison group despite less reported study time. This finding indicated that the treatments in the R+ and R++ conditions facilitated learning of the target vocabulary due to the vocabulary tasks, although participants required less out-of-class study time to achieve the same degree of vocabulary knowledge as those in the NR condition.

The third finding was that the treatments positively affected overall writing quality. Analysis of the tracking essays indicated that the target vocabulary was used correctly more frequently by participants in stronger treatment conditions than those in weaker conditions. In addition, the results also indicated that the developmental trajectories for lexical complexity and the number of standard length words for the participants in stronger treatment conditions was greater than those in weaker conditions.

The fourth finding was that the stronger treatments typically resulted in greater perceived usefulness of the vocabulary study. After one semester, during which the treatments were administered, the participants in stronger treatment conditions reported greater perceived usefulness of the vocabulary study as it pertained to their writing. After the second semester, when the treatments had ceased, the perceived usefulness of the vocabulary study became more positive across all groups indicating that it takes time for learners to appreciate the usefulness of tasks, such as the vocabulary learning tasks used in this study, in a language classroom.

Limitations

Five limitations need to be addressed. The main limitation is the generalizability of the findings of this study. Four factors might limit this. First, this study was conducted in natural conditions, so it was not possible to have a true randomization of participants nor was it possible to completely isolate participants in one group from another. Second, multiple instructors were involved in administering the reminders and providing additional instructions. Therefore, it was impossible to fully control how these reminders or instructions were given. A third factor was that the multiple pre- and posttests might have resulted in test fatigue, which in turn affected the results. Finally, there was no vocabulary testing before the start of the fall semester due to institutional constraints, which made it difficult to determine if any attrition of vocabulary occurred between the first and second semesters.

The second limitation involved the accuracy and completeness of the quantitative data. Much of the data were incomplete, hence these are unreliable for determining differences between groups. Although care was taken by the Vocabulary Course teachers to remind the participants to avail themselves of the Quizlet (2017) study sets and take the practice quizzes on Moodle (2016), the zero-stakes nature of Quizlet (2017) resulted in large numbers of students opting out of its use, resulting in gaps in the data. In addition, a potential weakness in the Quizlet and Moodle (2016) data were that these were measures of frequency and not duration, which, as a measure of time, would have provided a better estimate for the amount of vocabulary study. The problem with counting the number of attempts in Quizlet (2017) or Moodle (2016) potentially lies in participants not completing the tasks to the best of their ability and instead simply fulfilling course requirements. Measures of duration would show if the participants were taking the time needed to adequately complete the task or were just skimming through it.

Another source of potential inaccuracy was the fact that the self-study logs were susceptible to problems due to their reliance on the participants' memory. Previous studies have shown that relying solely on memory questionnaires is inadvisable as participants are not able to recall events perfectly and that collected data tends to be distorted (e.g., Hermann, 1982; Zuriff, 2003). With the reported amount of self-study was one measure in the study, this problem of reliability is of concern. The failure to use the Quizlet (2017) and Moodle (2016) activities suggests that curriculum designers and course planners need to be aware that many students ignore out-of-class study opportunities if they are optional.

Fatigue also likely affected the accuracy of the data. The vocabulary inventory survey data were collected to support the interpretation of other data. For each of the three measurements, the participants were given the full list of 500 target vocabulary items and asked to check off the words that they knew. This task likely resulted in fatigue, as several of the interviewees, who were considered motivated language learners, reported that they could not review the whole list with total accuracy. Regardless, while the data had some gaps, they provided information regarding the role of study time on vocabulary learning.

The third limitation was the consistency of the topics of the tracking essays. During the study, the writing prompt for the tracking essay asked the participants to write about a memorable event in their lives. This prompt was the same in all three administrations of the tracking essay. However, due to institutional constraints, a more precise prompt such as a memorable event that occurred during their high school was not permitted. This restriction resulted in topic inconsistencies between the three tracking essays participants produced. Only 32 participants (18%) wrote about the same topic in all three administrations of the tracking essay. As vocabulary used tends to be associated strongly with topic, a change in topic for the tracking essays would have affected the interpretation of the amount of learned and retained vocabulary.

The fourth limitation of the study was my lack of experience as an interviewer. During the interviews, there were several instances where, during the coding process, I became increasingly aware that I had failed to adequately probe the interviewee. As a result, some of the interviews lacked sufficient detail or depth. This missing data could have provided further insights into the participants' behaviors and beliefs. Retrospectively, I believe the questions concerning the target vocabulary identified by the participants' as being known or unknown could have followed the interview structure used by Nagy, Herman, and Anderson (1985). In this interview structure, the interviewees are asked to read the target vocabulary from flashcards. Next, the interviewee are asked to define the word or use it in a sentence. This process would reveal partial or complete knowledge of the target vocabulary.

The fifth limitation concerns the GTEC. The GTEC was used to measure the participants' language proficiency during the study and was already being used at the university prior to the start of the study. Replacing it with another test such as the TOEIC was not possible. Furthermore, in the context of this study, the GTEC was a zero-stakes test, which possibly influenced the results if test-takers were not fully committed to achieving a high score.

Suggestions for Future Research

The current study was a first step in considering transfer of knowledge in a language program, and the results suggest several directions for research to build upon the work reported in this study.

Although the current study indicated that there are secondary or peripheral effects of the treatments such as improvements in writing quality, a follow-up study might provide more accurate data by changing elements of the tracking essay. First, more precise wording of the writing prompt to narrow the range of possible topics would improve the comparability of the tracking essays and reduce the effect of topics on the use of vocabulary. Doing this would make it easier to evaluate the amount of learning and retention. Second, the tracking essays could be administered more frequently over a longer period. As vocabulary knowledge is incremental, this additional data might help researchers tease out changes in vocabulary knowledge.

Another potential area of study involves more closely monitoring participants' beliefs regarding the reminders, the vocabulary-focused tasks, the effects of these on language development, and the perceived usefulness of the vocabulary. This investigation would require a skilled interviewer and participants who are capable of describing their thoughts and feelings in detail. This purpose would be best achieved by conducting interviews more regularly over the course of the study. It would also be best achieved if done in the participants' L1.

Final Comments

Transfer of knowledge is the ultimate aim of learning and yet educators have little understanding of how to evaluate or influence this. As a teacher of English as a foreign language, I understand how various elements of a language program are connected to each other and that mastery of a certain skill in one course is intended to be applied in other courses. However, students at all levels of education tend to compartmentalize these skills and treat them as tasks to complete for one course. Application to the wider context has been assumed but rarely checked or evaluated.

Having completed a one-year project investigating transfer of vocabulary knowledge to written texts, it has become apparent that language learners can be more successful when made aware of how something they are learning can be used in other contexts. The use of tasks designed to raise this awareness and put into practice what they are learning can serve to reinforce knowledge and aid in transfer. While language learners will almost certainly improve with instruction, this improvement can be made faster and more permanent when they become aware of the need to use knowledge from one course in another.

Considerable discussion in curriculum design and pedagogy concerns students' failure to apply knowledge to wider contexts. Teachers wonder why students fail to use skills and knowledge despite having previously demonstrated competence. I hope that this study broadens the discussion to include reminders as a potential factor in the success or failure to transfer knowledge.

The main takeaway of this study is that reminders are a worthwhile consideration in curriculum design and pedagogy. Simple reminders had an effect on vocabulary learning, retention, and use. However, this effect was more pronounced when the reminders were paired with a noticing activity such as highlighting or deliberate word-selection. Furthermore, there were other benefits including greater perceived usefulness of vocabulary study, potential reduction in the amount of dedicated study time, and more rapid gains in overall language development and writing quality. In short, the optimal strength of reminder in this study was the reminder and highlighting task (R+ group). Participants who were in the R+ group showed markedly greater gains in vocabulary learning, retention, and use compared to the R group. While the R++ group did produce marginally greater improvement, the degree of improvement over the R+ group belied the need for this level of intervention.

Furthermore, the current study was inspired by my realization that the students were not expected to apply the knowledge they learned in one course to another course. While this is logical, as assessing students based on knowledge acquired elsewhere might seem inappropriate, for a program to ignore the transfer of knowledge between courses left me unsatisfied. Language skills are inherently applicable in multiple situations. For example, many language courses teach vocabulary with the underlying goal being that the learners will use the words in different contexts. However, vocabulary tests are typically administered in the course where the words are taught, as it cannot be reasonably expected that another course can evaluate this learning. There are a number of reasons why such evaluations are impractical or impossible.

In the current study, I set out to first establish whether the vocabulary that the students had been taught was being used by them after one and two semesters and whether this might indicate learning and retention of the target vocabulary. I was also interested in determining whether the strength of reminders affected vocabulary use, the correct use of the words, and the learners' study habits. Transfer of learning has been extensively discussed outside of SLA and I was interested in establishing the degree to which this transfer existed in the foreign language classroom. Through the study, I attempted to gain an understanding of students' views of their vocabulary learning, its usefulness as it relates to their writing, and the connection between courses and methods of assessment. The final aim was to determine if the influence of the reminders resulted in overall language improvement as operationalized by writing quality.

Overall, the findings regarding the effects of the reminders were mixed. The analysis of the amounts and correctness of target vocabulary usage showed that there were no significant differences between the groups. This conclusion is informed by Coxhead's (2008) findings, which suggested that usage of underlined or highlighted target vocabulary required at least partial word knowledge. In other words, the participants in the R+ and R++ groups might have identified potentially useful target vocabulary words but deemed these too risky for use due to insufficient knowledge of the terms. Similarly, the treatments did not have a statistically significant effect on the amount of study time. However, it was noted that the lack of significant findings might have been the result of several factors including a lack of sensitivity in the instruments and the incremental nature of vocabulary learning. While the quantitative results were largely inconclusive and require further research, the qualitative data showed more promise regarding the use of reminders as the participants reported an effect on writing quality and perceived the reminders to have been useful when they were salient. Furthermore, the reminders and vocabulary-focused tasks seem to have potential for reducing the amount of dedicated, out-of-class vocabulary study time due to the opportunity for this learning to take place during the task. Similarly, in determining the change in the participants' writing quality, the data suggests that the reminders might have affected the rates of improvement in overall writing scores, lexical complexity, and writing fluency. In addition, other studies have shown that learner-generated marking, which the R+ and R++ groups did, is as effective for assessments as multiple-choice tests (Leutner et al., 2007), free-recall tests (Rickards & August, 1975), short-answer tests (Amer, 1994; Blanchard & Mikkelson, 1987), fill-in-the-blank tests (Yue et al., 2015), and essay questions (Annis & Davis, 1978; Rickards & Denner, 1979).

Having established that the treatments might have had an effect on transfer of learning, it then becomes of interest to both teachers and researchers to better understand the potential impact of reminders on overall language development. The research questions concerned with correlational analyses attempted to illuminate the possible associations between study time, vocabulary tests, essays written under test conditions, and broader language proficiency measures. In essence, the line of inquiry used for this study was an attempt to explore a chain of events starting with the reminder treatments. The results suggest that the treatments translate into use of a larger number of target vocabulary, a higher incidence of correct use, greater target vocabulary learning and retention, more rapid gains on standardized language tests, and improvement in overall writing quality.

REFERENCES

Amer, A. A. (1994). The effect of knowledge-map and underling training on the reading comprehension of scientific texts. *English for Specific Purposes*, *13*, 35–45. https://doi.org/10.1016/0889-4906(94)90023-X

Anderson, R. C., Stevens, K. C., Shifrin, Z., & Osborn, J. (1978). Instantiation of word meanings in children. *Journal of Reading Behaviour*, *10*, 149–157. https://doi.org/10.1080/10862967809547264

Annis, L., & Davis, J. K. (1978). Study techniques and cognitive style: Their effect on recall and recognition. *The Journal of Educational Research*, *71*, 175–178. https://doi.org/10.1080/00220671.1978.10885064

Aoki, M. (2016, September 5). English heads for elementary school in 2020 but hurdles abound. *The Japan Times*. Retrieved from http://www.japantimes.co.jp/

Arevart, S., & Nation, I. S. P. (1991). Fluency improvement in a second language. *RELC Journal*, *22*(1), 84–94. https://doi.org/10.1177/003368829102200106

Aslin, R. N. (1993). Segmentation of fluent speech into words: Learning models and the role of maternal input. In D. Maurer (Ed.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 305–315). Dordrecht, Netherlands: Springer.

Astika, G. G. (1993). Analytical assessments of foreign students' writing. *RELC Journal*, *24*(1), 61–70. https://doi.org/10.1177/003368829302400104

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, *85*(3), 139–152. https://doi.org/10.1037/0033-295X.85.3.139

Baddeley, A. D. (1997). *Human memory: Theory and practice.* Bristol, England: Psychology Press.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*(5), 316–321. https://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *13*(2), 344–349. https://doi.org/10.1037/0278-7393.13.2.344

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press

Baumann, J. F. (2005). Vocabulary-comprehension relationships. In B. Maloch, J. V. Hoffman, D. L. Schallert, C. M. Fairbanks, & J. Worthy (Eds.), *Fifty-fourth Yearbook of the National Reading Conference* (pp. 117–131). Oak Creek, WI: National Reading Conference.

Benesse. (n.d.). *Shiboudai kitto mitsukaru kensaku* [Search for your preferred University]. Retrieved from http://kou.benesse.co.jp/hensachi/index.php

Benesse. (2004). GTEC: Global Test of English Communication. Retrieved from http://www.benesse.co.jp/gtec/english/

Benesse. (2017). *TOEIC TOEFL to no kansan te-buru* [TOEIC TOEFL conversion table]. Retrieved from http://www.benesse.co.jp/gtec/contact/answer/toeic_toefl_conversion.pdf

Blanchard, J., & Mikkelson, V. (1987). Underlining performance outcomes in expository text. *The Journal of Educational Research*, *80*, 197–201. https://doi.org/10.1080/00220671.1987.10885751

Bod, R. (1992). A computational model of language performance: Data oriented parsing. In A. Zampolli (Ed.), *Proceedings of the 14th conference on computational linguistics-Volume 3* (pp. 855–859). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/992383.992386

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Mahwah, NJ: Erlbaum.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*, 294–301. https://doi.org/10.1016/S1364-6613(99)01350-9

Broad, M. L., & Newstrom, J. W. (1992). Transfer of training: Action-packed strategies to ensure high payoff from training investments. Reading, MA: Addison-Wesley.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, *74*(1), 1–15. https://doi.org/10.1037/h0024206

Browne, C. (2002). To push or not to push: A vocabulary research question. *Aoyama Ronshu, 11*(02). Retrieved from: https://pdfs.semanticscholar.org/f10c/dd8afcbf29549cea6ad2b63da685e385c987.pdf

Browne, C., & Culligan, B. (2016). The TOEIC Service List. Retrieved from http://www.newgeneralservicelist.org

Browne, C., Culligan, B., & Phillips, J. (2013a). The New Academic Word List. Retrieved from http://www.newgeneralservicelist.org

Browne, C., Culligan, B., & Phillips, J. (2013b). The New General Service List. Retrieved from http://www.newgeneralservicelist.org

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal*, *10*, 12–42. Retrieved from http://jalt-publications.org/jj/

Bybee, J. (2003). 19 Mechanisms of change in grammaticization: The role of frequency. In B. Joseph & R. Janda (Eds.), *The handbook of historical linguistics* (pp. 602–623). Oxford, England: Blackwell.

Bybee, J. (2009). Language universals and usage-based theory. In M. H. Christiansen, C. Collins, & S. Edelman (Eds.), *Language universals* (pp. 17–39). New York, NY: Oxford University Press.

Byrnes, H., & Manchón, R. M. (Eds.). (2014). *Task-based language learning–Insights from and for L2 writing* (Vol. 7). Philadelphia, PA: Benjamins.

Carpenter, S.K., Cepeda, N.J., Rohrer, D., Kang, S.H.K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378. https://doi.org/10.1007/s10648-012-9205-z

Carroll, J. B. (1963). Research on teaching foreign languages. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1060–1100). Chicago, IL: Rand McNally.

Carver, R. P. (1976). Word length, prose difficulty, and reading rate. *Journal of Reading Behavior*, *8*(2), 193–203. https://doi.org/10.1080/10862967609547176

Carver, R. P. (1982). Optimal rate of reading prose. *Reading Research Quarterly*, *18*(1), 56–88. https://doi.org/10.2307/747538

Casanave, C. P. (2004). *Controversies in second language writing: Dilemmas and decisions in research and instruction*. Ann Arbor, MI: University of Michigan Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning*, *Memory and Cognition*, *19*, 389–396. https://doi.org/10.1037/0278-7393.19.2.389

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. *Computer Assisted Language Learning*, *18*(3), 231–251. https://doi.org/10.1080/09588220500185769

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268. https://doi.org/10.1080/016909698386528

Coady, J., Magoto, J., Hubbard, P., Graney, J., & Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. *Second Language Reading and Vocabulary Learning*, 217-228.

Cobb, T. (n.d.-a). Text lex compare version 3 [Computer software]. Retrieved from http://www.lextutor.ca/cgi-bin/tl_compare/index.pl

Cobb, T. (n.d.-b). Vocabprofiler [Computer software]. Retrieved from https://www.lextutor.ca/vp/

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, *12*(4), 345–360. https://doi.org/10.1076/call.12.4.345.5699

Coltheart, M. (1981). Disorders of reading and their implications for models of normal reading. *Visible Language*, *15*(3), 245–286. Retrieved from http://visiblelanguagejournal.com/

Corson, D. (1985). *The lexical bar*. Oxford, England: Pergamon Press.

Coxhead, A. (1998). *An academic word list.* Victoria University of Wellington, New Zealand.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2008). *Using vocabulary from input texts in writing tasks* (Unpublished doctoral dissertation). Victoria University of Wellington, New Zealand.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Creswell, J., & Plano Clark, V. (2010). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

De la Fuente, M. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition*, *24*, 81–112. https://doi.org/10.1017/S0272263102001043

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

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162–170. https://doi.org/10.1037/0022-0663.79.2.162

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In R. Bjork & E. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

Desse, J. (1958). Transfer of training: The psychology of learning. New York, NY: McGraw-Hill.

Disappointing levels of English [Editorial]. (2015, March 28). *The Japan Times.* Japan Times Online. Retrieved from http://www.japantimes.co.jp

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2d

Eckes, T. (2015). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.). Frankfurt am Main, Germany: Peter Lang.

Ekuni, R., Vaz, L. J., & Bueno, O. F. A. (2011). Levels of processing: The evolution of a framework. *Psychology & Neuroscience*, *4*(3), 333–339. https://doi.org/10.3922/j.psns.2011.3.006

Elgort, I. (2007). The role of intentional decontextualized learning in second language vocabulary acquisition: Evidence from primed lexical decision tasks with advanced bilinguals (Unpublished doctoral dissertation). Victoria University of Wellington, New Zealand.

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

Elias, C. S., & Perfetti, C. A. (1973). Encoding task and recognition memory: The importance of semantic encoding. *Journal of Experimental Psychology*, *99*(2), 151–156. https://doi.org/10.1037/h0034644

Ellis, H. C. (1965). *The transfer of learning.* New York, NY: Macmillan.

Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, *8*(2-3), 103–128. https://doi.org/10.1080/0958822940080202

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(2), 143–188. https://doi.org/10.1017/s0272263102002024

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44. https://www.doi.org/10.1017/S0267190512000025

Ellis, N. C. (2015). Cognitive and social aspects of learning from usage. In T. Cadierno & S. W. Eskildsen, *Usage-based perspectives on second language learning* (pp. 49–74). Berlin, Germany: Walter de Gruyter.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2015). Usage-based language learning. In B. MacWhinney, & W. O'Grady (Eds.), *The handbook of language emergence* (1st ed.) (pp. 49–75). Chichester, England; John Wiley & Sons.

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams, *Theories in second language acquisition: An introduction* (pp. 75–93). New York, NY: Routledge.

Ellis, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, *21*, 85–301. https://doi.org/10.1017/S0272263199002077

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. https://doi.org/10.1207/s15516709cog1402_1

Eskildsen, S. W. (2008). Constructing another language—Usage-based linguistics in second language acquisition. *Applied Linguistics*, *30*(3), 335–357. https://doi.org/10.1093/applin/amn037

Educational Testing Services [ETS]. (2016). Test and data score summary for TOEFL iBT tests: January 2015-December 2015 test data. Retrieved from https://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Educational Testing Services [ETS]. (2017). Test of English for international communication (TOEIC). Retrieved from https://www.ets.org/toeic

Educational Testing Services [ETS]. (2018). 2018 Report on test takers worldwide: The TOEIC listening and reading test. Retrieved from https://www.ets.org/s/toeic/pdf/2018-report-on-test-takers-worldwide.pdf

Educational Testing Service [ETS]. (2019). TOEFL iBT public use dataset. TOEFL committee of examiners 2020 research program. Retrieved from https://www.ets.org/toefl/grants/coe_research_program

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157–169. https://doi.org/10.1111/j.2044-8295.1978.tb01643.x

Field, A. (2013). Discovering statistics using IBM SPSS statistics (4th ed.). London, England: Sage.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fisher, W. P., Jr. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238. Retrieved from https://www.rasch.org/rmt/ rmt63i.htm

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095. Retrieved from https://www.rasch.org/rmt/ rmt211m.htm

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal*, *36*(2), 239–248. https://doi.org/10.1080/09571730802390759

Fletcher, P. C., Shallice, T., Frith, C. D., Frackowiak, R. S., & Dolan, R. J. (1998). The functional roles of prefrontal cortex in episodic memory. II. Retrieval. *Brain*, *121*(7), 1249–1256. https://doi.org/10.1093/brain/121.7.1249

Fogarty, R., Perkins, D., & Barell, J. (1992). *How to teach for transfer: The mindful school*. Highett, Australia: Hawker Brownlow Education.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. https://doi.org/10.2307/40264523

Friederici, A. D., Opitz, B., & Von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral Cortex*, *10*(7), 698–705. https://doi.org/10.1093/cercor/10.7.698

Gagne, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning*. New York, NY: HarperCollins College Publishers.

Gick, M., & Holyoak, K. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–42). New York, NY: Academic Press.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*, 95–112. https://doi.org/10.3758/BF03197590

Gries, S. Th., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning, 65*(1), 228–255. https://www.doi.org/10.1111/lang.12119

Griffin, G., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, *17*(4), 443–460. https://doi.org/10.1017/S0142716400008195

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Hackman, J. R., & Oldham, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley.

Hatch, J. A. (2002). *Doing qualitative research in education settings*. New York, NY: State University of New York Press.

Hatch, J. A. (2002). *Doing qualitative research in education settings*. New York, NY: State University of New York Press.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303–317. https://doi.org/10.1017/s0272263199002089

Herrmann, D. J. (1982). Know thy memory: The use of questionnaires to assess and study memory. *Psychological Bulletin*, *92*, 434–452. https://doi.org/82/9202-0434S00.75

Higa, M. (1965). The psycholinguistic concept of "difficulty" and the teaching of foreign language vocabulary. *Language Learning*, *15*(3–4), 167–179. https://doi.org/10.1111/j.1467-1770.1965.tb00799.x

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal*, *36*(2), 153–166. https://doi.org/10.1080/09571730802389983

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–97). Potomac, MD: Erlbaum.

Hopper, P. J. (1998). Emergent grammar. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (Vol. 1, pp. 155–176). New York, NY: Psychology Press.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning & Technology*, *9*(2), 90–110. Retrieved from http://llt.msu.edu/

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal, 41*, 301–312. https://doi.org/10.1177/0033688210380558

Hu, H. C. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System*, *56*, 28–39. https://doi.org/10.1016/j.system.2015.11.001

Huang, C. S., Yang, S. J., Chiang, T. H., & Su, A. Y. (2016). Effects of situated mobile learning approach on learning motivation and performance of EFL students. *Journal of Educational Technology & Society*, *19*(1), 263–276

Hughey, J. B., Wormuth, D. R., Hartfiel, V. F., & Jacobs, H. L. (1983). *Teaching ESL composition: Principles and techniques*. Rowley, MA: Newbury House.

Hulstijn, J. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, England: Cambridge University Press.

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/10.1111/0023-8333.00164

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 471–480. https://doi.org/10.1016/S0022-5371(73)80027-1

Ishii, T. (2015). Semantic connection or visual connection: Investigating the true source of confusion. *Language Teaching Research, 19*, 712–722. https://doi.org/10.1177/1362168814559799

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition.* Rowley, MA: Newbury House.

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, *19*(3), 357–377. https://doi.org/10.1093/applin/19.3.357

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.) *Formulaic sequences* (pp. 269–291). Amsterdam: Benjamins.

Jordan, R. R. (1997). Identification of problems and needs: a student project. In A.P. Cowie & J.B. Heaton (Eds.), English for academic purposes (pp. 67–89). Reading, UK: University of Reading.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. *Language Learning*, *16*(1-2), 1–20. https://doi.org/10.1111/j.1467-1770.1966.tb00804.x

Kapur, S., Craik, F. I., Tulving, E., Wilson, A. A., Houle, S., & Brown, G. M. (1994). Neuroanatomical correlates of encoding in episodic memory: Levels of processing effect. *Proceedings of the National Academy of Sciences*, *91*(6), 2008–2011. https://doi.org/10.1073/pnas.91.6.2008

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Katona, G. (1940). *Organizing and memorizing*. New York, NY: Columbia University Press.

Kaur, J., & Hegelheimer, V. (2005). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, *18*(4), 287–310. https://doi.org/10.1080/09588220500280412

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 365–386. https://doi.org/10.1177/1362168808089922

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28–44. Retrieved from http://llt.msu.edu/

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kim, Y., Lowenstein, K. L., Pearson, P. D., & McLellan, M. (2001). A framework for conceptualizing and evaluating the contributions of written language activity in oral language development for ESL students. In J. Hoffman, D. Schallert, C. Fairbanks, J. Worthy, & B. Maloch (Eds.), *50th yearbook of the National Reading Conference* (pp. 333–345). Chicago, IL: National Reading Conference.

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals, 22*(6), 529-540. https://doi.org/10.1111/j.1944-9720.1989.tb02780.x

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits-and costs-of dropping flashcards. *Memory*, *16*, 125–136. https://doi.org/10.1080/09658210701763899

Kuru, Y., Masaki, M., & Kinshi, K. (2009). The preliminary examinations of the rubric of our own designing. *The Bulletin of the Writing Research Group*, *JACET Kansai Chapter*, *8*, 27–40.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC journal, 25*(2), 21–33. https://doi.org/10.1177/003368829402500202

Laufer, B. (1997). The lexical plight in second language reading. In Coady, J. and T. Huckin (Eds.), *Second Language Vocabulary Acquisition*, (pp. 20–34). Cambridge, England: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Laufer, B. (2003a). The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression. In V. Cook (Ed.), *Effects of the second language on the first*, (pp. 19–31). Clevedon, England: Multilingual Matters.

Laufer, B. (2003b). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, *59*(4), 567–587. https://doi.org/10.3138/cmlr.59.4.567

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108. https://doi.org/10.1177/003368829702800106

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, *31*(4), 537–561. https://doi.org/10.1016/j.system.2003.02.004

Leeser, M. J. (2008). Pushed output, noticing, and development of past tense morphology in content-based instruction. *Canadian Modern Language Review*, *65*(2), 195–220. https://doi.org/10.3138/cmlr.65.2.195

Leutner, D., Leopold, C., & Den Elzen-Rump, V. (2007). Selfregulated learning with a text-highlighting strategy: A training experiment. *Zeitschrift für Psychologie/Journal of Psychology*, *215*, 174–182. https://doi.org/ 10.1027/0044-3409.215.3.174

Lewis, M. (1993). The lexical approach: The state of ELT and a way forward. Hove, England: Language Teaching.

Linacre, J. M. (n.d.-a). Reliability and separation measures. Retrieved from http://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (n.d.-b). Fit diagnosis: infit outfit mean-square standardized. Retrieved from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. Retrieved from https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-Model Computer Program*. Chicago, IL: Winsteps.com. Retrieved from http://ifile.hkedcity.net/1/001/950/public/Secondary/EI0020070012/winsteps.pdf

Linacre, J. M. (2013). *WINSTEPS: Rasch measurement computer program* (version 3.80.0) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2014). *FACETS: Many-facet Rasch measurement computer program* (version 3.71.4) [Computer software]. Chicago, IL: Winsteps.com.

Lockhart, R. S., & Craik, F. I. (1978). Levels of processing: A reply to Eysenck. *British Journal of Psychology*, *69*(2), 171–175. https://doi.org/10.1111/j.2044-8295.1978.tb01644.x

Lockhart, R. S., & Craik, F. I. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *44*(1), 87–112. https://doi.org/10.1037/h0084237

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15*, 257–280. https://doi.org/10.1080/13825580701322171

Malone, J. (2018). Incidental vocabulary learning in SLA: effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, *40*(3), 651–675. https://doi.org/10.1017/S0272263117000341

Marmurek, H. H. (1995). Encoding, retrieval, main effects and interactions: Were Lockhart and Craik (1990) on the level? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *49*(2), 174–192. https://doi.org/10.1037/1196-1961.49.2.174

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for specific purposes, 28*(3), 183–198. https://doi.org/10.1016/j.esp.2009.04.003

Maurice, K. (1983). The fluency workshop. *TESOL Newsletter*, *17*(4), 130. (ERIC Document Reproduction Service No. ED289347).

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 47–55. Retrieved from http://vli-journal.org/wp/

Meara, P. (2009). Connected words: Word associations and second language vocabulary acquisition. Amsterdam, The Netherlands: Benjamins.

Meara, P. M., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size test, Version E1.1/K10*. Zurich: Eurocentres Learning Service.

Meleis, A. I., & Dagenais, F. (1980). Response Bias and Self-Report of Honesty. *The Journal of General Psychology, 103*(2), 303-304. https://doi.org/10.1080/00221309.1980.9921010

Memrise Limited. (n.d.). Memrise [Mobile & on-line computer application]. Retrieved from https://www.memrise.com/

Meyers, P. C. (2013). Generation. *ELT Journal*, *67*(2), 230–232. https://doi.org/10.1093/elt/ccs063

Miller, G. A., Bruner, J. S., & Postman, L. (1954). Familiarity of letter sequences and tachistoscopic identification. *The Journal of General Psychology*, *50*(1), 129–139. https://doi.org/10.1080/00221309.1954.9710109

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American*, *257*(3), 94–104. https://doi.org/10.1038/scientificamerican0987-94

Ministry of Education, Culture, Sports, Science and Technology [MEXT]. (n.d.). Overview. Retrieved from http://www.mext.go.jp/en/policy/education/overview/index.htm

Ministry of Education, Culture, Sports, Science and Technology [MEXT]. (2008). Shogakkou gakushu shidou yoryo eigoban gaikokugo katsudou [Study of course guideline for foreign language activities in elementary schools; provisional version]. Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/1261037.htm

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*(3), 390–407. https://doi.org/10.1177/1745691617710510

Mizuno, R. (2003). The application of cognitive scientific findings related to spacing effects to web-based CAI systems. In *Proceedings of the Joint Workshop of Cognition and Learning through Media-Communication for Advanced e-Learning* (pp. 128–133). Berlin, Germany: Japanese-German Center Berlin.

Mondria, J. A., & Mondria-de Vries, S. (1994). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. *System*, *22*(1), 47–57. https://doi.org/10.1016/0346-251X(94)90039-6

Moodle Pty. (2016). Moodle [Learning Management System]. Retrieved from https://moodle.org/

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Murai, S. (2016, January 25). Changes in store for TOEIC, but test still not total gauge of fluency: Experts. *The Japan Times*. Retrieved from http://www.japantimes.co.jp/

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233–253. https://doi.org/10.2307/747758

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(01), 3–20. https://doi.org/10.1017/S0958344008000219

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. https://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning. *Studies in Second Language Acquisition, 37*, 677–711. https://doi.org/10.1017/S0272263114000825

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition, 41,* 287–311. https://doi.org/10.1017/S0272263118000219

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (1980). Strategies for receptive vocabulary learning. *Guidelines*, *3*(1), 18–23. Retrieved from https://www.wgtn.ac.nz/lals/about/staff/publications/ paul-nation/1980-Receptive-vocab.pdf

Nation, I. S. P. (1982). Beginning to learn foreign vocabulary: A review of the research. *RELC Journal*, *13*(1), 14–36. https://doi.org/10.1177/003368828201300102

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Newbury House.

Nation, I. S. P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal, 9*(2), 6–10.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description*, *acquisition and pedagogy* (pp. 6–19). Cambridge, England: Cambridge University Press.

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

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(2), 151–171. https://doi.org/10.1016/S0022-5371(77)80044-3

Nesi, H., & Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. *System 22*, 1–15. https://doi.org/10.1016/0346-251X(94)90036-1

Newfields, T. (2006). Suggested answers for assessment literacy self-study quiz #1. *Shiken*, *10* (2), 25–32. Retrieved from http://jalt.org/test/SSA1.htm

Nyberg, L. (2002). Levels of processing: A view from functional brain imaging. *Memory*, *10*(5–6), 345–348. https://doi.org/10.1080/09658210244000171

Ong, J., & Zhang, L. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218–233. https://doi.org/10.1016/j.jslw.2010.10.003

Osborne, R., & Wittrock, M. (1985). The generative learning model and its implications for science education. *Studies in Science Education*, *12*(1), 59–87. https://doi.org/10.1080/03057268508559923

Paltridge, B., & Starfield, S. (Eds.). (2013). *The handbook of English for specific purposes* (Vol. 592). West-Sussex, England: Wiley-Blackwell.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. Neville Postlethwaite (Eds.), *International encyclopedia of education*, *2* (pp. 6452–6457). Oxford, England: Pergamon Press.

Perkins, D. N., & Salomon, G. (1996). Learning transfer. In A. Tuijnman (Ed.), *International encyclopedia of adult education and training* (pp. 422–427). Kidlington, England: Elsevier Science.

Piaget, J. (1926). *The language and thought of the child.* London, England: Kegan Paul, Trench, Trubner and Company.

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, *51*(2), 73–75. https://doi.org/10.1111/j.1540-4781.1967.tb06700.x

Quizlet Inc. (2017). Quizlet [Mobile app and computer software]. Retrieved from https://quizlet.com/

Random.org. (2019, January). Random integer set generator. Retrieved from https://www.random.org/integer-sets/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedogogiske Instiut. https://doi.org/10.1016/S0019-9958(61)80061-2

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*, 77–89. https://doi.org/10.2307/3585941

Rickards, J. P., & August, G. J. (1975). Generative underlining strategies in prose recall. *Journal of Educational Psychology*, *67*, 860–865. https://doi.org/10.1037/0022-0663.67.6.860

Rickards, J. P., & Denner, P. R. (1979). Depressive effects of underlining and adjunct questions on children's recall of text. *Instructional Science*, *8*, 81–90. https://doi.org/10.1007/BF00054983

Ripple, R. E., & Drinkwater, D. J. (1982). Transfer of learning. In H. Mitzel, J. Best, & W. Rabinowitz (Eds.), *Encyclopedia of educational research* (pp. 1947–1953). New York, NY: Free Press.

Royer, J. M. (1973). Memory effects for test-like-events during acquisition of foreign language vocabulary. *Psychological Reports*, *32*(1), 195–198. https://doi.org/10.2466/pr0.1973.32.1.195

Sandberg, J., Maris, M., & Hoogendoorn, P. (2014). The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning. *Computers & Education*, *76*, 119–130. https://doi.org/10.1016/j.compedu.2014.03.006

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69–90. https://doi.org/10.2307/3587062

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description*, *acquisition and pedagogy* (pp. 199–227). Cambridge, England: Cambridge University Press.

Schmitt, N. (2007). Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (pp. 827–842). Norwell, MA: Springer.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.

Schmitt, N. & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal*, *49*(2), 133–143. https://doi.org/10.1093/elt/49.2.133

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1191/026553201668475857

Schneider, M. L., & Fujishima, N. K. (1995). When practice doesn't make perfect: The case of a graduate ESL student. In D.Belcher, & G. Braine (Eds.), *Academic writing in a second language: Essays on research & pedagogy* (pp. 231–265). Norwood, NJ: Ablex.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603. https://doi.org/10.1126/science.275.5306.1599

Simcock, M. (1993). Developing productive vocabulary using the "Ask and Answer" Technique. *Guidelines*, *15*(2), 1–7. (ERIC Document Reproduction Service No. EJ485094).

Sorrell, C. J. (2012). Zipf's law and vocabulary. In C. A. Chapelle (Ed.), *Encyclopaedia of Applied Linguistics*. Oxford, England: Wiley-Blackwell.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110. https://doi.org/10.3102/00346543056001072

Stevens, V. (1995). Concordancing with language learners: Why? When? What? *CAELL*, *6*(2), 2–9. Retrieved from http://www.ritsumei.ac.jp/~nozawa/CAELLJ.html

Stewart, D. (2016, October). Japan gets schooled: Why the country's universities are failing. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/articles/japan/2016-10-31/japan-gets-schooled

Stoeckel, T., & Bennett, P. (2015). New general service list test (NGSLT). Retrieved from http://www.newgeneralservicelist.org/ngsl-levels-test/

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38*, 244–253. https://doi.org/10.3758/MC.38.2.244

Sugimoto, Y. (2010). *An introduction to Japanese society*. Cambridge, England: Cambridge University Press.

Sun, Y. C. (2007). Learner perceptions of a concordancing tool for academic writing. *Computer Assisted Language Learning*, *20*(4), 323–343. https://doi.org/10.1080/09588220701745791

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. Language Learning, 65(4), 860-895. https://doi.org/10.1111/lang.12138

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 165–179). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of HG Widdowson*, *2*(3) (pp. 125–144). Oxford, England: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). Cambridge, England: Cambridge University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 1, pp. 471–483). New York, NY: Routledge.

Tahira, M. (2012). Behind MEXT's new course of study guidelines. *The Language Teacher*, *36*(3), 3–8. Retrieved from http://jalt-publications.org/tlt

Thorndike, E. L. (1908). Memory for paired associates. *Psychological Review*, *15*(2), 122–138. https://doi.org/10.1037/h0073570

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions: III. Functions involving attention, observation and discrimination. *Psychological Review*, *8*(6), 553–564. https://doi.org/10.1037/h0071363

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1/2), 61–82. Retrieved from http://www.cognitivelinguistics.org/en/journal

Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *14*, 145–152. https://doi.org/10.1037/0278-7393.14.1.145

Tulving, E. (1979). Relation between encoding specificity and levels of processing. In L. Cermak & F. Craik (Eds.), *Levels of processing in human memory* (pp. 405–428). Mahwah, NJ: Erlbaum.

Tulving, E. (2001). Episodic memory and common sense: How far apart? In A. Baddeley, M. Conway, J. Aggleton, J. Hodges, & K. Graham (Eds.), *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *356*(1413) (pp. 1505–1515). London, England: Royal Society Publishing.

Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*(1), 64–75. https://doi.org/10.1080/09571736.2016.1191527

Wakabayashi, S. (2015, August 17). What is lacking in English education in Japan? [Editorial]. *Chuo Online.* Yomiuri Shimbun Advertising Bureau. Retrieved from https://yab.yomiuri.co.jp/adv/chuo/dy/

Wang, Y. H., & Shih, S. K. H. (2015). Mobile-assisted language learning: Effects on EFL vocabulary learning. *International Journal of Mobile Communications*, *13*(4), 358–375. https://doi.org/10.1504/IJMC.2015.070060

Waring, R. (2004). In defence of learning words in word pairs: But only when doing it the ‘right’ way! Retrieved from http://www.robwaring.org/vocab/principles/systematic_learning.htm

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163. Retrieved from http://nflrc.hawaii.edu/rfl/

Webb, S. A. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. A. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*(3), 441–470. https://doi.org/10.3138/cmlr.65.3.441

Webb, S. A. & Chang, A. C-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Webb, S. A., Newton, J., & Chang, A. C-S. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Webb, S., & Rodgers, M. P. (2009). Vocabulary demands of television programs. *Language Learning, 59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Wertheimer, M. (1959). *Productive thinking*. New York, NY: Harper & Row.

West, M. (1953). *A general service list of English words*. London, England: Longman, Green & Co.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Wilkins, D. A. (1972). *Linguistics in language teaching*. Cambridge, MA: MIT Press.

Wilkinson, D. (2017). EFL Vocabulary Acquisition through Word Cards: Student Perceptions and Strategies. *TESL-EJ, 21*(3), n3.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579. https://doi.org/10.1080/09658211.2012.687052

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

Wong, A. (2017). Changes in students' L2 vocabulary-learning beliefs and strategies. *Hirao School of Management Review*, *7*, 21–53. Retrieved from https://konan-u.repo.nii.ac.jp/

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, *17*(1), 905–906. Retrieved from https://www.rasch.org/rmt/rmt171a.htm

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370–371. Retrieved from http://rasch.org/rmt/rmt83b.htm

Wulff, S., & Ellis, N. C. (2008). Statistical measures for usage ‐ based linguistics. In B. VanPatten, & J. Williams, *Theories in second language acquisition: An introduction* (pp. 75–93). New York, NY; Routledge.

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review*, *27*, 69–78. https://doi.org/10.1007/s10648-014-9277-z

Zipf, G. (1949). Human behavior and the principle of least effort: An introduction to human ecology. New York, NY: Hafner.

Zou, D. & Xie, H. (2018). Personalized word-learning based on technique feature analysis and learning analytics. Educational Technology & Society: Journal of International Forum of Educational Technology & Society and IEEE Learning Technology Task Force, 21(2), 233–244. Retrieved from https://www.jstor.org/stable/26388402

Zuriff, G. E. (2003). A method for measuring student study time and preliminary results. *College Student Journal*, *37*(1), 72–79. (Gale Document No. A99816483).

APPENDICES

APPENDIX A

PILOT STUDY 1 INFORMED CONSENT FORM (ENGLISH)

Vocabulary Use in Writing
Principal Investigator: Aeric Wong
April, 2017

I would like to analyze your writing assignments to further understand how students use vocabulary in writing. As this might lead to a research paper, I would like to ask your permission to collect your writing assignments for use in my research.

Benefits:
There are no direct benefits to you from taking part in this study.

Risks:
There are no risks to you from taking part in this study. Neither participation in this nor lack of participation will influence your grade.

Confidentiality:
Every effort will be made to maintain the privacy of your data. No one will know who provided the data used in the study.

Study Withdrawal:
You have the right to withhold your consent. Deciding not to participate will have no effect on benefits you receive from the course you are enrolled in.

Invitation for Questions:
If you have questions about this study, contact the Principal Investigator, Aeric Wong, before you sign this consent form. If you have questions at a later date, please contact him by e-mail:

*****@********.ac.jp

If you have no objections, please sign the form below in both places and return the bottom portion. If you would like to be receive a copy of any paper that might be derived from these data, please add your e-mail address to the bottom form.

Figure A1. English version of the informed consent form for Pilot Study 1, page 1.

Copy for participants

Authorization:

I have read this paper about the study or it was read to me. I know the possible risks and benefits. I know that being in this study is voluntary. I understand the above and I choose to participate in the study. I know that I can withdraw at any time. I received a copy of this document on the date that is shown.

Name of Participant (printed) ____

Signature of Participant ____________________________ Date _________________

Vocabulary Use in Academic Writing
Principal Investigator: Aeric Wong
October, 2015

Copy for researcher

Authorization:

I have read this paper about the study or it was read to me. I know the possible risks and benefits. I know that being in this study is voluntary. Understanding the above, I choose to participate in this study. I know that I can withdraw at any time. I, on the date signed, received a copy of this document.

Name of Participant (printed) ____

Signature of Participant ____________________________ Date _________________

Contact e-mail (*optional: provide to receive a copy of any research papers*):

____

Figure A1. English version of the informed consent form for Pilot Study 1, page 2.

APPENDIX B

PILOT STUDY 1 INFORMED CONSENT FORM (JAPANESE)

大学でのライティングにおける、ボキャブラリーの使用
調査責任者：エリック　ウォング
2015 年 10 月

私は、"大学でのライティングにおいて、学生がどのようにボキャブラリーを使うか"という、より一層深い理解のため、あなたのライティング課題を解析したいと思います。この解析が、今後の研究文書につながる可能性があるので、あなたのライティング課題をデータとして集めることに対して、同意頂けるかどうかをお伺いをします。

メリット:
この研究調査に参加することで、あなたに直接的なメリットは発生しません。
デメリット:
この研究調査に参加することへのデメリットも発生しません。また、参加しないことによって、成績に影響することはありません。
守秘義務:
あなたのデータのプライバシーを守るよう、最善を尽くします。この研究で使用されるデータは、誰から提供されたかどうか、分からないようになります。
参加撤回:
あなたには、この同意を取り下げる権利があります。参加しないと決めたことにより、登録している授業で受けるメリットに影響を及ぼすことはありません。
質問がある場合:
この研究に関して質問がある場合は、この同意書に署名する前に調査責任者である、エリック ウォングにお知らせください。後日、質問が出た場合は、下記のメールでご連絡下さい。awong@center.konan-u.ac.jp

もし、参加に同意して頂ける場合は、下記の 2 つの箱内両方に署名し、下の部分を返却して下さい。
このデータが使用される可能性ある研究文書のコピーを受け取り希望であれば、下の部分のメールアドレス欄に記入して下さい。

Figure B1. Japanese version of the informed consent form for Pilot Study 1, page 1.

参加者控え:

承諾:

私は、研究について、本書を読みました、もしくは第三者が読み上げてくれました。メリット、デメリットについて理解しました。また研究への参加が任意・無償のものであることも理解しました。その上で参加することを選択しました。いつでも参加辞退可能である旨も理解しました。私は、日付が記載された、本書のコピーを受け取りました。

参加者名 (printed) ____

参加者署名 ________________________________ 日付____________________

大学でのライティングにおける、ボキャブラリーの使用

調査責任者:エリック　ウォング

2015 年 10 月

調査者控え:

承諾:

私は、研究について、本書を読みました、もしくは第三者が読み上げてくれました。メリット、デメリットについて理解しました。また研究への参加が任意・無償のものであることも理解しました。その上で参加することを選択しました。いつでも参加辞退可能である旨も理解しました。私は、日付が記載された、本書のコピーを受け取りました。

参加者名 (printed) ____

参加者署名 ________________________________ 日付____________________

メールアドレス (研究文書のコピーを受け取り希望の場合のみ記入):

____

Figure B1. Japanese version of the informed consent form for Pilot Study 1, page 2.

APPENDIX C

EXAMPLES OF VOCABULARY TEST SECTIONS

Vocabulary Test 4 (NGSL 1001-1200)

Section 1: Matching - Match the words in each group on the left with their meanings on the right. Write the correct number in the space provided.

A.	1) context (n.)	____ an area that is distinct from other areas
	2) majority (n.)	
	3) alternative (n.)	____ the circumstances and setting for an event or idea
	4) sector (n.)	
	5) extent (n.)	____ the bigger number
	6) audience (n.)	

B.	1) handle (v.)	____ to direct, lead or guide something
	2) trust (v.)	
	3) conduct (v.)	____ to watch carefully; to make a comment
	4) inform (v.)	
	5) observe (v.)	____ to have the ability to cope with a situation or problem
	6) feed (v.)	

C.	1) appropriate (adj.)	____ being right or suitable for some situation or purpose
	2) primary (adj.)	
	3) federal (adj.)	____ system of government
	4) surface (adj.)	
	5) responsible (adj.)	____ first or of most importance
	6) spot (adj.)	

D.	1) slightly (adv.)	____ being separated by an amount of time or space
	2) apart (adv.)	
	3) directly (adv.)	____ completely; totally; very
	4) otherwise (adv.)	
	5) absolutely (adv.)	____ in the end
	6) eventually (adv.)	

E.	1) element (n.)	____ the position or rank relative to others in a society
	2) administration (n.)	
	3) attitude (n.)	____ a part of something
	4) principle (n.)	
	5) status (n.)	____ the way you act, think, and feel about something
	6) category (n.)	

Figure C1. Sample vocabulary test used in the vocabulary course, page 1.

Section 2: Multiple Choice - *Circle the word that best matches the definition.*

1. He was forced to _____ that he didn't try hard enough.
   a. examine b. admit c. transfer
2. A person's behaviour can _____ a lot about their character.
   a. reveal b. spread c. observe
3. She wasn't _____ that there were classes this weekend.
   a. responsible b. aware c. mistake
4. There are many companies that can _____ letters and packages all over the world.
   a. launch b. tour c. deliver
5. It is interesting to see how technology has changed over the past _____.
   a. category b. print c. decade
6. Despite efforts by the government, the birth rate continues to _____.
   a. cry b. flow c. decline
7. They didn't _____ my request for an interview.
   a. grant b. trust c. ensure
8. He was known as a very _____ man.
   a. fair b. overall c. primary
9. She didn't make a single mistake. Everything was _____.
   a. cross b. fix c. perfect
10. Many people don't like reading a _____ that is too long.
   a. text b. band c. collection

Section 3: Original Sentences - Choose 5 words from the box below. Use each of the words in a separate sentence. You can change the forms of the words. For example, for *run* you could use *ran*, or for *energy* you could use *energize*, or *energetically*. Underline the word that you are using.

request	exchange	correct	collect	progress
launch	respond	struggle	consumer	warn

1. ____

____

2. ____

____

3. ____

____

4. ____

____

5. ____

____

Figure C1. Sample vocabulary test used in the vocabulary course, page 2.

APPENDIX D

PILOT STUDY 2 INFORMED CONSENT FORM (ENGLISH)

INFORMED CONSENT FORM

Title of Study Japanese university students' vocabulary learning experiences, beliefs, and strategy use and the effect of guided instruction

Researcher Aeric Wong, Ph. D. student in Applied Linguistics at Temple University Japan

Purpose of the Form:

The purpose of this form is to ask you to participate in a research study. I will explain the study to you and ask you to volunteer to participate in the study. Please feel free to ask any questions that you may have before you decide if you want to participate. Whether you take part in the study is entirely up to you, and you are also free to withdraw from the study at any time. By signing this consent form, you are not waiving any legal rights that you have. You are only indicating that the study has been explained to you, that you have had an opportunity to ask questions, that you understand what is involved, and that you are willing to participate in the study.

Explanation of the study

The goal of this study is to investigate the motivations of English language learners regarding vocabulary study in their courses and how it pertains to reading and writing tasks in their discipline courses. Of particular interest is how the students perceive the vocabulary study, the strategies employed during the year, if they have an increased awareness of the target vocabulary in their content-based classes, and if their study habits and motivations have changed over the course of the academic year. A secondary purpose of this study is for me to gain experience with qualitative research methods. This project is part of my course work at Temple University Japan, but it may evolve into a larger study to be published in the future. Data collected in the study will be used for educational and research purposes only.

Imposition on participants

The estimated duration of your participation in the study is 1 hour. During this time, I would like to interview you about your vocabulary learning in the CUBE English class. I would also like to record the audio of the interview. In total, participation in this study is estimated to involve 1 hour of your time.

Figure D1. English version of the informed consent form for Pilot Study 2, page 1.

Risks

The reasonably foreseeable risks or discomforts of participating in this study are minimal. However, if at any time, you do not feel comfortable answering a question, having your interaction recorded, or having me observe you, please let me know and I will respect your requests.

Benefits

The benefit of participating in this study is knowing that you have contributed to an increased understanding of vocabulary acquisition and to the education of a future researcher.

Your rights as a research participant

Your participation is voluntary. You may decide not to participate and you are free to stop participating in the research at any time without penalty. If you decide not to participate at any time, no information related to you will be included in this study.

At any time now and in the future, you may ask me questions about this study and I will do my best to answer your questions to your satisfaction.

Confidentiality

I will do my best to keep the sources of the data collected for this study confidential. I will keep all of this information in a safe place and do my best to ensure your anonymity. All the proper names (names of people, institutions, places, etc.) used in this study, including yours, will be changed. I will also do my best to limit the disclosure of your identity, but I cannot promise complete secrecy. There is always a potential risk of loss of confidentiality, and there is a slight chance that fellow researchers who hear your recorded voice or see videotaped data might recognize you. In addition, in order to ensure that I am following the rules and regulations regarding research and the protection of human subjects, some organizations (e.g., the IRB at Temple University) may inspect and copy your information.

Dissemination

Data collected for this project will go into assignments for my coursework in the Intermediate Qualitative Research course at Temple University Japan. I will also be using the data in a class presentation to my fellow classmates. In the future, the data may also be used for a publication.

Figure D1. English version of the informed consent form for Pilot Study 2, page 2.

Your consent to participate in this study

Please initial the ways, if any, in which you agree to participate in this study.

1. I do not want to participate in this study in any way. ______
2. I agree to participate in this research. ______
3. I agree to be audio-recorded during the interview. ______
4. I also give permission to Aeric Wong to retain the audio-recordings for 6 months for educational and research purposes only. ______
5. I give permission to Aeric Wong to quote me using a pseudonym in future published paper/s. ______

I understand the information given to me, and I have received answers to any questions I have about the study. I understand and agree to the conditions of this study.

If you have any other conditions regarding the use of the data provided for this study, please specify them below.

....

....

I understand that my participation in this study is voluntary, and that I may withdraw from this study at any time by telling Aeric Wong to remove me from the study. Contact information is written below.

I also understand that I will receive a signed copy of this consent form.

____

Name (please print)

______________________________ ____________________

Signature Date

Thank you very much for your participation.

Aeric Wong

______________________________ ____________________

Signature Date

Figure D1. English version of the informed consent form for Pilot Study 2, page 3.

APPENDIX E

PILOT STUDY 2 LANGUAGE LEARNING AND VOCABULARY SURVEY

(ENGLISH)

1) How long have you been studying English? ___________ years.

2) Did you go to a juku? Yes / No

2a) If yes, when did you go (circle all that apply, write the number of hours each week)

a. Elementary school (_____ hours per week)

b. Junior High School (_____ hours per week)

c. High School (_____ hours per week)

3) Please circle all the places you studied English and write how many years you studied in each.

a. Elementary school (_____ years)

b. Junior High School (_____ years)

c. High School (_____ years)

d. Eikaiwa (_____ years)

e. Juku (_____ years)

f. Home tutor (_____ years)

g. Other (please specify: ____________________) (_____ years)

4) Do you study English outside university classes now? Yes / No

If yes, which do you do? (circle all that apply)

a. Talk with friends

b. Talk with family

c. Listen to English music

d. Watch English movies

e. Send messages to friends (e.g., on Line, Facebook, etc.)

f. Other : ____

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 1.

5) Please circle all the things you studied in junior/senior high school

a. Vocabulary
b. Grammar
c. Writing short sentences
d. Writing paragraphs
e. Writing essays
f. Listening
g. Speaking – Pronunciation
h. Speaking – Conversation
i. Reading short story books
j. Reading stories/passages in textbooks
k. Test preparation (Eiken, TOEIC, TOEFL, etc.)
l. Other (please specify: ____)

6) Please circle all the things you studied in JUKU

a. Vocabulary
b. Grammar
c. Writing sentences
d. Writing paragraphs
e. Writing essays
f. Listening
g. Speaking – Pronunciation
h. Speaking – Communication
i. Reading short story books
j. Reading stories/passages in textbooks
k. Test preparation (Eiken, TOEIC, TOEFL, etc.)
l. Other (please specify: ____)

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 2.

7) Why did you study English (please circle all that apply)

a. My teacher told me to
b. My parents told me to
c. To pass tests/exams
d. I want to speak with English-speakers
e. I want to watch movies in English
f. I want to listen to music in English
g. I want to travel to English-speaking countries
h. Other (please specify: ____)

8) Have you traveled/lived overseas?

No Yes

Where? ____

How long? ____

9) How did you study English vocabulary? (Please circle all that apply)

a. Write the Japanese translation of words
b. Write explanations in Japanese
c. Write the English meaning
d. Write example sentences in Japanese
e. Write example sentences in English
f. Use word cards
g. Use smartphone application
h. Use website
i. Use computer program
j. Use dictionaries
k. Complete vocabulary questions in textbooks
l. Memorize English vocabulary by writing their spelling
m. Reading books
n. Talking with people

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 3.

10) When you studied vocabulary, where did the words come from?

a. My teacher made the word list

b. A test preparation book (e.g., TOEIC questions)

c. General English practice book (e.g., grammar exercise books not for examinations)

d. I chose the words

11) When did you improve your English the most?

a. Elementary School

b. Junior High School

c. High School

11a) Why do you think you improved at that time?

a. Study for entrance exams

b. Study for the TOEIC/TOEFL/IELTS

c. Personal interest (more information: ___________________________________)

12) What do you think students should study in English class? (Please circle all that apply)

a. Vocabulary

b. Grammar

c. Writing essays

d. Writing personal stories

e. Listening practice

f. Reading short stories

g. Pronunciation

h. Speaking

i. Practicing for tests (Eiken, TOEIC, TOEFL, etc.)

j. Spelling

k. Other: ________________________

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 4.

13) In your own life, where do you use English now?

a. In university English classes

b. In university management classes

c. Talking with teachers

d. Talking with friends

e. Conversation practice with friends

f. Doing homework for English classes

g. Doing homework for management classes

h. Clubs or Circles (e.g., ESS)

i. Test practice (TOEIC, TOEFL, IELTS, etc.)

j. I don't use English

14) Rank the English skills

(Reading, Writing, Speaking, Listening)

Most important ________________________________

____

____

Least important ________________________________

14a) Why is ___ the most important?

15) I think studying English will help me in my job after I graduate (Circle as applicable)

I do not particularly think so ------------- I very much think so

0 1 2 3 4

16) I think studying English will help me in my social/personal life after I graduate (Circle as applicable)

I do not particularly think so ------------- I very much think so

0 1 2 3 4

17) I think studying English will help me in my (non-English) classes at university (Circle as applicable)

I do not particularly think so ------------- I very much think so

0 1 2 3 4

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 5.

18) Please describe how you study vocabulary now

19) How did you learn the study style you wrote in question 18? (school teacher, school friends, brother/sister, etc.)

20) How can you improve the study style you wrote in question 18? What improvement is needed?

21) What words do you think are important to study?

a. For business

b. For daily conversation

c. For taking tests

d. For subjects in university

e. Other (____)

22) In Japan, you have many chances to use English

I do not particularly think so ------------- I very much think so

0 1 2 3 4

23) Outside of English class, where do you think you will use English now or in the future?

24) I think English is useful Yes / No

If yes, how is it useful?

25) On average, how many times do you think you need to meet (read or hear it) a NEW word to remember the meaning?

26) Which is the most difficult part of language to learn?

Figure E1. English version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 6.

APPENDIX F

PILOT STUDY 2 LANGUAGE LEARNING & VOCABULARY SURVEY

(JAPANESE)

1) 英語をどれくらいの期間勉強していますか？ ___________ 年

2) 学習塾に通いましたか？　はい／いいえ

2a)　2)が「はい」の場合、いつ行きましたか？

(該当するもの全てに○し、1 週間あたりの学習時間を記入して下さい。)

a. 小学校　(_____時間／週)

b. 中学校　(_____時間／週)

c. 高校　(_____時間／週)

3) 英語を勉強した場所全てに○をし、それぞれ何年間勉強したか記入して下さい。

a. 小学校　(_____年間)

b. 中学校　(_____年間)

c. 高校　(_____年間)

d. 英会話学校　(_____年間)

e. 塾　(_____年間)

f. 家庭教師　(_____年間)

g. その他 (具体的な場所: ______________________________) (_____年間)

4) 大学の授業外で英語をいま勉強していますか？　はい／いいえ

「はい」の場合、下記のどれをしていますか？ (該当するもの全てに○をして下さい。)

a. 友達と話す

b. 家族と話す

c. 英語の音楽を聴く

d. 英語の映画を観る

e. メッセージを友達に送る (例：LINE や Facebook など)

f. その他 : ____

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 1.

5) <u>中学校・高校</u>で勉強したもの全てに○をしてください。

a. ボキャブラリー
b. 文法
c. 短文を書く
d. パラグラフを書く
e. エッセーを書く
f. リスニング
g. スピーキング- 発音
h. スピーキング- 会話
i. 短い物語の本を読む
j. テキストの物語/文章を読む
k. テスト対策 (英検,TOEIC, TOEFL 等)
l. その他(具体的に: ____)

6) <u>塾</u>で勉強したもの全てに○をしてください。

a. ボキャブラリー
b. 文法
c. 短文を書く
d. パラグラフを書く
e. エッセーを書く
f. リスニング
g. スピーキング- 発音
h. スピーキング- 会話
i. 短い物語の本を読む
j. テキストの物語/文章を読む
k. テスト対策（英検,TOEIC,TOEFL 等）
l. その他(具体的に：____)

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 2.

7) なぜ英語を勉強しましたか？(該当するもの全てに〇をして下さい。)
a. 学校の先生が勉強するように言ったのから
b. 両親が勉強するように言ったのから
c. テスト／試験に合格するため
d. 英語を話す人たちと話したいから
e. 英語で映画を観たいから
f. 英語の音楽を聴きたいから
g. 英語圏を旅行したいから
h. その他 (具体的に：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿)

8) これまでに海外を旅行した／住んだことはありますか？
いいえ　／　はい
a. 場所 ＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿
b. 期間 ＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

9) 英語のボキャブラリーをどのように勉強しましたか？(該当するもの全てに〇をして下さい。)

a. 日本語の（直訳の）単語を書く
b. 日本語の詳細な意味（説明）を書く
c. 英語の意味を書く
d. 日本語の例文を書く
e. 英語の例文を書く
f. 単語帳を使う
g. スマートフォンアプリを使う
h. ウェブサイトを使う
i. コンピュータプログラムを使う
j. 辞書を使う
k. テキストでボキャブラリーの問題を解く
l. 英単語のスペルを書いて覚える
m. 本を読む
n. 人と話す

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 3.

10) ボキャブラリーを勉強した際、単語はどこから来たものでしたか？

a. 先生が単語リストを作った

b. 試験用問題集（例：TOEIC 問題集）

c. 一般的な英語学習用問題集（例：試験用ではない一般的な文法問題集）

d. 自分で単語を選択した

11) あなたが一番英語が上達したのはいつですか？

a. 小学校

b. 中学校

c. 高校

11a) なぜ 11)の時に上達したと思いますか？

a. 入試の為に勉強したから

b. TOEIC/TOEFL/IELTS のために勉強したから

c. 個人的な興味 (具体的に: ____)

12) 英語の授業で、学生は何を勉強すべきだと思いますか？(該当するもの全てに〇をして下さい。)

a. ボキャブラリー

b. 文法

c. エッセーを書く

d. 個人的な話を書く

e. リスニングの練習

f. 短い物語を読む

g. 発音

h. スピーキング

i. テスト勉強 (TOEIC / TOEFL 等)

j. スペル

k. その他: ____________________

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 4.

13) 生活の中で、いまどこで英語を使いますか？
a. 大学の英語の授業
b. 大学のマネージメントの授業
c. 先生と話す
d. 友達と話す
e. 友達と会話の練習をする
f. 英語の授業の宿題をする
g. マネージメントの授業の宿題をする
h. クラブまたはサークル(例：ESS)
i. テスト勉強 (TOEIC,TOEFL, IELTS, など)
j. 英語は使いません

14) 下記の英語スキルを重要な順に挙げてください。
(リーディング、ライティング、スピーキング、リスニング)
最も重要 ______________________________
____
____
最も重要でない ______________________________

14a) 14)で「最も重要」と答えたものを選んだ理由は何ですか？

15) 英語の勉強は、卒業後の仕事を探すのに役立つと思う。（該当するものに○）
特にそう思わない -----------------すごくそう思う
0 1 2 3 4

16) 英語の勉強は、卒業後の社会生活/個人生活に役立つと思う。（該当するものに○）
特にそう思わない -----------------すごくそう思う
0 1 2 3 4

17) 英語の勉強は、[ ][1] での英語以外の授業に役立つと思う。（該当するものに○）
特にそう思わない -----------------すごくそう思う
0 1 2 3 4

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 5.

[1] *Name identifying university department site removed.*

18) 今現在、どのようにボキャブラリーを勉強しているか書いてください。

19) 18)で答えた勉強のスタイルをどのようにして得ましたか／知りましたか？
(例：学校の先生や学校の友達、兄弟/姉妹から教えてもらったなど)

20) 18)で答えた勉強のスタイルは完ぺきなスタイルですか？改善の余地はありますか？どのような改善が必要ですか？

21) どのような単語を勉強するのが大切だと思いますか？
a. ビジネス用
b. 日常会話用
c. テスト受験用
d. 大学での授業用
e. その他____)

22) 日本で英語を使うチャンスがたくさんある
特にそう思わない -----------------すごくそう思う
0 1 2 3 4

23) 英語の授業外で、いまもしくは将来的にあなたはどこで英語を使うと思いますか？

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 6.

24) 英語は役に立つと思う　はい／いいえ
「はい」と答えた場合、どのよう（なこと）に役立つと思いますか？

25) 新しい単語の意味を覚えるには、平均して何回その単語を読んだり聞いたりする必要があると思いますか？

26) 言語を学習するうえで、どんな部分が一番難しいと思いますか？

Figure F1. Japanese version of the Language Learning & Vocabulary Survey for Pilot Study 2, page 7.

APPENDIX G

PILOT STUDY 2 SEMI-STRUCTURED INTERVIEW QUESTIONS

General Learning History

1. How many years have you been studying English?
2. Did you study English in Elementary school/JHS/HS/eikaiwa/juku?
   a) Outside of school, how many hours a week did you study English? EL, JHS, HS?
3. Did you study for English Entrance Exams in elementary school, Junior high school, high school? If so, how did you study? What part, if any at all, did vocabulary learning play in your studies for entrance exams? *Note: Were there English Entrance Exams? How did you study? Did you study vocabulary?*
4. When do you think you made the most progress in learning vocabulary? *Note: Level up*
5. Have you had any overseas experience? *Note: Travel*, *study*, *homestay*, *etc.*
6. What did you study in English class in school?

   (vocabulary, *grammar*, *writing sentences*, *paragraphs*, *essays*, *listening*, *speaking*, *pronunciation*, *communication*, *reading short story books*, *reading stories / passages in textbooks*, *test preparation)*

7. What did you study outside of English class?
8. In your own life, where do you use English now? (University English classes, talking with teachers / friends, conversation practice with friends, doing English homework, clubs / circles, test practice)

VL History and Strategies

9. How, if at all, did you study English **vocabulary** in elementary school? Junior high school? High school? Juku? Eikaiwa? (word cards, direct translation, sentence construction, etc.)
10. What VL strategies were you taught by teachers, friends, parents, etc.? What rationale, if any, was given for these approaches to VL?
11. Which of these did you find useful? Which do you still use now?
12. When you studied vocabulary? What would you study?

    Note: Meaning (Japanese=> English? English=> Japanese?)? Pronunciation? Part of speech? Word in sentence? Spelling? Etc.
13. How long have you been using this method of vocabulary study?
14. How would you choose which words to study? Words from lists? *Tangocho* 単語帳? Readings in textbooks?
15. Why did you think you were studying these words?

16. What are your thoughts of this method? (Good? Bad? Adequate? Suitable for you?)

Beliefs

17. Which did you think was most useful for you?

18. Why did you study English? (was it self-driven or teacher/parent-directed?)

    a) For teacher/parent direction, why do you think they wanted you to study English?

19. Which English skills do you think are the most useful? (Writing, Speaking, Listening, Reading)

20. Why should Japanese university students study English?

21. How, it at all, do you think English will help you in your social life?

22. How, if at all, do you think English will help you in your professional/working life?

Questions on VL Beliefs?

23. Now, what do you think is the best method for studying vocabulary?

24. Which vocabulary words do you think are most important to study?

25. What aspects of words are important to study?

APPENDIX H

PILOT STUDY 2 SAMPLE INTERVIEW QUESTIONS BASED ON SURVEYS

April: GTEC L/R/T (113/91/**204**)

July: TOEIC L/R/T (190/140/**330**)

January: GTEC L/R/T (104/97/**201**)

Vocab Pre-Post: (50/49)

Italicized = Responses in post-survey

Underlined = Responses in pre-survey that did not appear in post-survey

1. What was your English learning experience since you were a child?
2. Is there a special reason you chose to come to this university?
3. In the survey, you said you practice English by talking with friends, and sending messages. What do you talk about? How often and how long?
4. Practice is good. Do you think it is helping you with your English ability? How about the classes?
5. Reason for studying English is for travel. Where? Why?
6. Did you use English in Romania?
7. What do you think students should study in English class? Why?

   Vocabulary, *Writing personal stories*, *Reading stories*, *Pronunciation*, *Speaking*

<u>Grammar, Writing essays, Listening practice</u>

No test prep?

8. For vocabulary, how do you study?

   Write L1 word, *L1 sentence*, *L2 sentence*, *Smartphone app*, *Textbook*, *Spelling*, *Reading*, *Talking*

   <u>L2 meaning, Dictionary use</u>

   Did your study style change from HS? Is there a way you can improve your study style now?

9. Where do you use English now?

   <u>Talking with friends, Conversation practice</u>

   Talking with teachers, *Class*, *Homework*

10. What is the most important skill?

    1 – Speaking

    Writing from 4 to 2

11. What kind of job would you like to do after you graduate? How will English help you?

12. How about your social life?

13. You said your vocabulary study before was reading books or speaking or messaging people. How did this help you learn vocabulary? How about now? Reading books and TOEFL textbook. Why the TOEFL book?
14. Where do you think you will use English outside class? Daily life. How much?
15. How many times do you need to meet a new word to remember the meaning?
16. What is the most difficult part to language to learn?
17. How do you think you will study vocabulary next semester? Do you think you will continue to use Quizlet?

APPENDIX I

PILOT STUDY 2 SAMPLE CODED SURVEY AND INTERVIEW SUMMARIES

In the transcript sections below, *Beliefs* are indicated by *italics*, Strategies are indicated by a double underline, **Changes** are indicated in **bold**.

Mami (a pseudonym)

Survey

- Messaging
- Watching CNN – Learned from a Japanese professor (TOEFL words). Take notes, compare with website.
- Students don't need test prep in class
- Only using English in university English class (no more managements class or talking with teachers)
- *Most important skill is* **reading** *(4-1) & speaking (1-2*)
    - o *Communication is important (more chance or options when job-hunting)*
    - o *Reading is for reading foreign documents (work focus)*
- HS [Write words, read meanings, write sentences] PLACE [**Quizlet**, Write, Read] Convenience of being able to be used on the train.

- Where do you think you will use English? *Work (Travel or service industry)*
- English is useful*? Job hunting*, *work*, **XTraveling**
- Most difficult part of language? **XRemember**, *Speaking* + *Listening*

Interview

- Seattle exchange program in JHS 2. 2 weeks. Still email friends from there (once a week). *Speaking is more important than writing.*
- Chances to use English: *English class*, *social events with foreigners.*
- Reading is important now for work. Reading documents. Wants to work in travel or service industry. Therefore, to prepare for this, what she should study is reading practice.
- Speaking is important for work 174 "for my future at work using English." Listening Speed is difficult.
- Wants to study vocabulary more 147-162. Business and daily conversation.
- Speaking & Listening is difficult. Ability to explain, Ability to understand.

APPENDIX J

PILOT STUDY 2 INDUCTIVE THEMES IDENTIFIED IN INTERVIEWS

Examples of inductive themes identified in interview responses for Pilot Study 2.

Speaking is more important

Conversation with "Mami" (lines 58 to 67)

M Speaking is more important than writing. So I want to learn speaking more because I want to speak English with foreigners.

I This answer I hear from many people, to communicate with people from foreign countries. I think it's a nice goal to have. What chances do you think you have to speak with foreign people?

M I want to go to foreign countries.

I So it's basically when you travel.

M Yes.

I Do you think you have chances in Japan?

M Umm … English class, and self-access center … I want to join with foreign parties for speaking time. For example, PKN.

Excerpts from conversation with "Sayuri" (line numbers indicated)

57 I want to talk to more fluently. So I want to practice speaking more. I don't like textbook speech. I like daily conversation.

61 The feedback from the GTEC is that my vocabulary is low. And my score went down. I talk with Eri about English and study abroad so Eri's advice is to learn more vocabulary. My vocabulary is not enough.

110 I think daily conversation is the most important. Business and university subjects … fly away. If I can talk daily conversation more smoothly, university subjects and business … I can speak.

Excerpt from conversation with "Nana" (line number indicated)

172 If I can speak English, I can talk not only with (native) English speakers, if they live in Europe, but maybe they use English. So, I went to Romania, Romanians can speak English. So people who can speak English, even easy English, I can communicate.

<u>Reading is most important (line 125 to 128)</u>

I It was interesting for me to see how your answer changed. In September, you said speaking was the most important, but in December, you said reading was the most important.

Mami Because now I learned in class, reading more books and when I work in a company, maybe I will read a lot of English documents, maybe.

<u>Use of Quizlet</u>

Mami (line 93 to 97)

M Using Quizlet is very good for me to study vocabulary. But many students don't use Quizlet, I think. So, students should use Quizlet more.

I Why is Quizlet so good for you?

M When I go to university, I can study on the train or when I can use PC. It's easy for me to study using Quizlet on the train.

Sayuri (line 7 to 19)

I I remember your first interview, you said you used to study by writing on paper and putting on the wall. Do you still do this?

S Mmm … no. Now, I don't do it.

I Oh, why not?

S Writing and putting on the wall takes a lot of time. Quizlet is not. So I use Quizlet now.

S Quizlet is … I have to collect the words I don't know. So I put them into Quizlet. This is troublesome at first but after the cards are made, the cards are always there. Putting words on the wall is also troublesome but it is a little troublesome all the time. Quizlet is very troublesome at first.

Nana (line 132 to 134)

N Both … my feeling … if I want to learn one vocabulary, I'm looking for myself. I'm looking for English sentence of that vocabulary. In Quizlet, other students find the information. That is not easy for me to remember.

<u>Strategies changing</u>

Mami (line 102-106)

M So you're using Quizlet on the train but you also do this? Are you only using Quizlet? Or are you using Quizlet and …

M I write English vocabulary and Japanese meaning.

I So it's your old style + Quizlet.

M Yes. This study style, I learned in high school so I keep using it now.

(line 201 to 206)

I How do you think you will study next semester?

M Now, even for the tests in CUBE English class, if I don't study so much, I'm still okay. But for the TOEIC test, the vocabulary is difficult so I need to spend more time studying those words. More than now.

I Okay. I know that the Quizlet app, many students were using it because I told them to use it. But next semester, the teachers won't say "you must use Quizlet." Do you think your study style will be different?

M Mmm … I want to use Quizlet. It's useful. I think maybe there's an app to study for the TOEIC so I will search.

Sayuri (line 31 to 33)

S Not the same … the words that are difficult to remember, I will put on the wall. Not so difficult words, I will use Quizlet. I can see many times on the wall. Quizlet is, I bring my phone every day but I don't use it so much. So I think difficult words will be on the wall.

<u>Strategies not changing</u>

Shoko (line 190 to 194)

Sh I didn't use Quizlet.

I That's okay. Is there a reason why you didn't use it?

Sh Writing on paper is easier for me to remember.

I Okay. So this is your own study style. Maybe Quizlet didn't match your style?

Sh Yeah. I prefer analog method.

<u>What vocabulary should we study?</u>

Mami (line 142 to 152)

I Okay. So for vocabulary, because there are many different types of vocabulary, what vocabulary do you think you should learn?

M Now I think, the vocabulary in class … there aren’t a lot of words so I buy a vocabulary word list book, and I should learn TOEIC words.

I So this is a TOEIC book?

M Yes. TOEIC vocabulary word list book. I should study this.

I No problem. I get the feeling that everything you’re doing, your English study, the words you want to study …

M I think that in other English classes, like [content course 1] or [content course 2], we can’t study vocabulary so much.

I Ah, okay. The vocabulary focus is not so strong.

M Yeah. So I think vocabulary study only in [the vocabulary course] is not enough.

Sayuri (line numbers indicated)

57 I want to talk to more fluently. So I want to practice speaking more. I don’t like textbook speech. I like daily conversation.

110 I think daily conversation is the most important. Business and university subjects … fly away. If I can talk daily conversation more smoothly, university subjects and business … I can speak.

<u>Cannot express myself. Words don't come out.</u>

Mami (line 173 to 195)

I Okay, this question, I want to find out, why is speaking the most difficult?

M Mmm … now I can't speak English well. So I should speak English more and more. Because for my future at work using English.

I Why do you think speaking is the most difficult? You didn't say listening is the most difficult or …

M I can't understand the speaking level of foreigners.

I Are the words difficult? Or is it the speed?

M Speed! Yes, speed. When I watch CNN news, sometimes I cannot understand the speaking speed. Very fast. And I want to explain well. When I want to explain many things in English but I can't speak much. So speaking is difficult, to explain things.

I So how … explaining is difficult. How can this become easier for you?

M Using English more every day.

I So it's practice.

M And I think I should go to foreign countries, exchange program, in the university.

I Okay. Do you think learning vocabulary will help you with this? With explaining?

M Learning vocabulary?

I Yeah. You said it is difficult to explain things. Will vocabulary help you with this?

M Yes. Speaking English … the words don't come out. So vocabulary skill … should more increase.

I Okay.

M When I write vocabulary, I can write, but I can't speak.

I Okay. You have time to think when you are writing …

M Yes! But when I speak, the words don't come out and I don't know how to explain.

I All right.

M So now I think speaking is difficult for me.

Sayuri (line 66)

S Now, I want to talk to you more casually but the words don't come out.

Shoko (line 62)

Sh Hard to tell in English what I think in Japanese. I use English phone dictionary, google translate …

Ayako (line 124 to 129)

A So I want to read faster. And I'm also not good at speaking. I can't say what my feel. So I need more vocabulary. So I should practice speaking

I So for the speaking part, of course we need vocabulary to communicate. What kinds of vocabulary do you think you need?

A When I'm just talking, I need daily vocabulary. So first of all, I need daily vocabulary. Difficult vocabulary is no need now.

Tomoko (line 199 to 201)

T Speaking.

I Why is it difficult?

T I don't have any vocabulary. And I can't explain what I'm thinking.

APPENDIX K

QUIZLET VOCABULARY SET SAMPLE

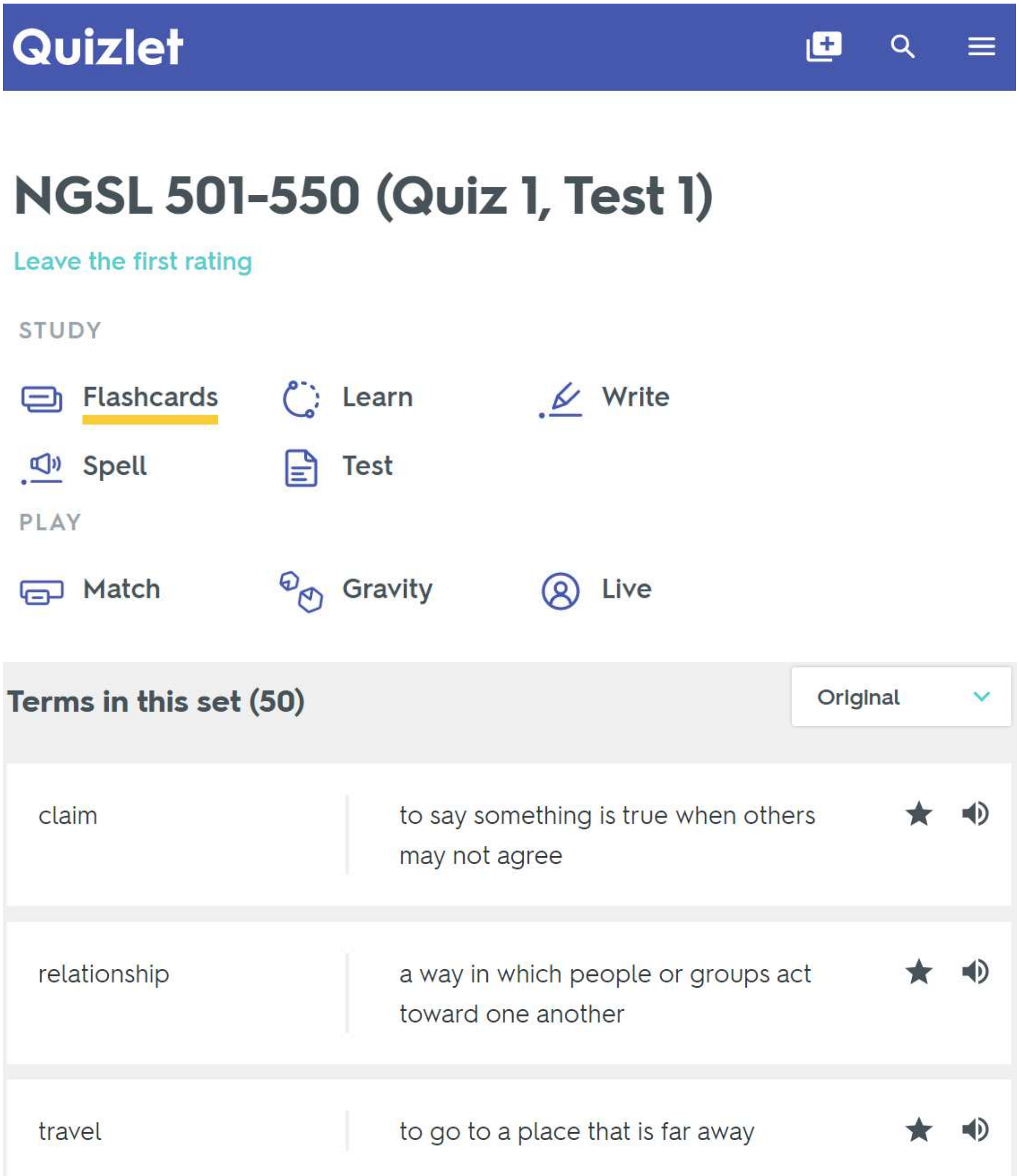

Figure K1. *Screen capture of a Quizlet (2017) test page for the NGSL 501-550 (Quiz 1, Test 1) set.*

APPENDIX L

GENRE WRITING TEACHING MATERIALS

Title

Background information

Events in chronological order

Paragraphs divided by major events

Evaluation with a comment

Past tense action verbs

Time phrases

3rd person voice

Adjectives

Steve Jobs

Steve Jobs was one of the *most important* people in *consumer* technology and created several of the *most common personal* electronics we use today. However, his life had a series of ups and downs, which helped him become a *great* entrepreneur.

Steve Jobs was born in San Francisco, California in 1955. **He** was adopted by Paul and Clara Jobs. **He** was a *good* student and became *good* friends with another student, Steve Wozniak. **They** began to build *simple* computers with other students. Jobs finished high school in 1972 and spent a *short* time in college before **he** dropped out and attended *various college* classes for the next 18 months. **He** then took a job at Atari in 1974 with Wozniak where **he** worked on circuit boards for video games. Jobs and Wozniak then started their own company, the Apple Computer Company the next year.

In 1976 **they** created the Apple I computer and received *commercial* success with the Macintosh computer in 1984. The company then grew *larger* but Jobs left the company in 1985 after arguments with other *company* leaders. **He** then purchased Pixar Animation studios for 10 million dollars the year after that, which became one of the most *successful* animation studios with Jobs as the CEO.

Jobs became CEO of Apple Computers again in the year 2000 and released the iPod player and iTunes software which became very *successful*. In 2006, **he** sold Pixar to Disney and received 7% of Disney stock. The next year, the iPhone was released, which made smartphones popular around the world. Jobs left Apple as CEO in 2011 due to *pancreatic* cancer and died six weeks later.

From his hard work and vision, he helped build one of the most famous companies around the world and created products that are loved by millions of people.

Figure L1. Example of a recount used in the content courses.

Recount - Framework

Guided Note-Taking - Steve Jobs

Steve Jobs was one most influential people in consumer technology and popularized several of the most common personal electronics we use today. However, his life had a series of ups and downs, which helped him become a great entrepreneur.

Steve Jobs was born in San Francisco in _____. He was a good student and graduated high school in ______ and spent a short time in college before dropping out and attending various classes for _____________. He then took a job at Atari in _____ where he worked on circuit boards for video games. Jobs then founded his own company, the Apple Computer Company __________.

In ______ he invented the Apple I computer and achieved success with the Macintosh computer in ______. However, Jobs resigned from the company in ______ after disagreements with other company leaders. He then purchased Pixar Animation studios for 10 million dollars ________, which became one of the most successful animation studios with Jobs as the CEO.

Jobs became CEO of Apple Computers again in ______ and released the iPod player which became wildly successful. In _______, he sold Pixar to Disney and received _____ of Disney stock. ________, the iPhone was released, which popularized smartphones around the world.

Jobs resigned as CEO of Apple in ______ due to severe pancreatic cancer and died six weeks later. His personal wealth at the time of his death was around __________________ and his name and company is known around the world.

Figure L2. Example of Recount Framework guided note-taking activity used in the content courses.

Why do we write recounts?

The purpose of a recount is to tell what happened. This may involve the author's personal interpretation of the events.

When do we write recounts?

Biographies, personal stories (journals), factual recounts (newspaper articles), imaginative recounts (eg. a day in the life of a puppy).

How do we write recounts?

1. Background
    a. Who? / Where? / When? / Why?
2. Events in time order
    a. First, → Finally,
    b. Details of events (May require a few paragraphs)
3. Evaluation
    a. Wrap up of the story, may include summary, include thoughts about the person

What language features do we see in a recount?

- Use of linking words related to time (when / after / during / before / next)
- Use of action verbs in the past tense (went, saw, returned, laughed)
- Use of personal pronouns in 3rd person (he, she, they)
- Use of adjectives to add detail to the story

Figure L3. Example of recount explanation sheet used for teaching this genre used in the content courses.

Famous American Basic Information

Name: ______________________________ Student #: _______________

1. Name: ____

2. Date of Birth: ____

3. Hometown: ____

4. Early Childhood events:

a. ____
____

b. ____
____

c. ____
____

5. Major Life events:

a. ____
____

b. ____
____

c. ____
____

d. ____
____

e. ____
____

6. Contribution to field:

____

____

____

7. Your thoughts:

____

____

____

Figure L4. Famous American Basic Information worksheet used in content courses.

Genre Writing - Recount - Rough Draft **Student Name: ________________________ Student #: _______________** **Directions**: Write a recount that gives details about the life of a famous American. Highlight the major actions and events of that person's life in chronological order.	
Background	
Events in order • **Use time connecting words** • **Write in the past tense** • **Use 3rd person active voice** • **Use adjectives**	

Figure L5. Sample of paper used for writing assignments in the content courses.

Recount Peer Check Sheet

Writer Name: ______________________ Reviewer Name: _________________

What are we looking for in a Recount (Biography)	1	2	3
Background (with information about who, what, where, when and why)			
Series of events in chronological (time) order			
Evaluation <u>with comment</u> on person's contribution			
Time Connectives (after ~, when, then, next)			
Details (adjectives/adverbs) to add interest			
Correct use of 3rd person voice			
Action verbs in the past tense			
Formatted appropriately (using clear paragraphs)			
Writing is clear and has a logical progression of ideas			
Writing is free of grammatical errors and spelling mistakes			

Notes / Comments **Total _____/50**

Reviewer Name: _____________________________

What are we looking for in a Recount (Biography)	1	2	3
Background (with information about who, what, where, when and why)			
Series of events in chronological (time) order			
Evaluation <u>with comment</u> on person's contribution			
Time Connectives (after ~, when, then, next)			
Details (adjectives/adverbs) to add interest			
Correct use of 3rd person voice			
Action verbs in the past tense			
Formatted appropriately (using clear paragraphs)			
Writing is clear and has a logical progression of ideas			
Writing is free of grammatical errors and spelling mistakes			

Notes / Comments **Total _____/50**

Figure L6. Recount Peer Check Sheet used in the content courses.

Student name:________________________ Student number: _____________

Writing Assignment 1 - Recount (Biography)

Directions: Retell the life story of a famous American in your own words. Include an introduction about the person and show why he or she is interesting. Then describe important episodes in their life in time order. At the end, give a conclusion telling your reader how important the person was and summarize the events you outline.

Name of famous American:________________________________

Figure L7. Page 1 of an assignment writing sheet used in the content courses. Page 2 contains additional lines for writing.

APPENDIX M

INFORMED CONSENT FORM (ENGLISH)

Figure D1 provides the English version of the informed consent form used for this study. The Japanese version of the informed consent form (see Appendix B) was attached to all students' papers when they wrote the Tracking Essay (see Appendix H) required by the program of study for placement in first-year English courses.

Writing by University Students
Principal Investigator: Aeric Wong
April, 2017

I would like to analyze your writing assignments to further understand students' writing. As this might lead to publication of the results, I would like to ask your permission to collect your writing, reports on your performance for exams, and other work already collected by the university for use in my research.

Benefits: There are no direct benefits to you from taking part in this study.
Risks: There are no risks to you from taking pa1t in this study. Neither participation nor lack of participation will influence your grade.
Confidentiality: Every effort will be made to maintain the privacy of your data. No one will know who provided the data used in the study.
Study Withdrawal: You have the right to withhold your consent. Deciding not to participate will have no effect on benefits you receive from the course you are enrolled in.
Invitation for Questions: If you have questions about this study, please contact the Principal Investigator, Aeric Wong, by e-mail before signing. If you have questions at a later date, please contact the Principal Investigator at the e-mail address below:
******@*******.ac.jp

If you do not want to participate, please sign. ________________________

Figure M1. English version of the informed consent form used in this study.
Note. Participants could decline to participate in the study or withdraw their data at any time. However, they still had to complete the course requirements, which included all the activities used for the study.

APPENDIX N

INFORMED CONSENT FORM (JAPANESE)

The Japanese version of the informed consent form used for this study appears in Figure E1. See Figure D1 (Appendix D) for an English translation.

大学生のライティング
研究責任者エリックウォング
2017 年 4 月

私は、学生のライティングをより深く理解するため、あなたのライティング課題を解析したいと思っています。この解析が今後の研究文書につながる可能性があるため、あなたのライティングや、試験結果レポート、また大学が既に回収したその他の課題などを私の研究に使用することに対して許可して頂けるかをお伺いします。

メリット: この研究調査に参加することで、あなたに直接的なメリットは発生しません。
デメリット: この研究調査に参加することへのデメリットも発生しません。また、参加しないことによって、成績に影響することはありません。
守秘義務: あなたのデータのプライバシーを守るよう、最善を尽くします。この研究で使用されるデータは、誰から提供されたかどうか、分からないようになります。
参加撤回: あなたには、この同意を取り下げる権利があります。参加しないと決めたことにより、登録している授業で受けるメリットに影響を及ぼすことはありません。
質問がある場合: この研究に関して質問がある場合は、この同意書に署名する前に調査責任者である、エリック ウォングにお知らせください。後日、質問が出た場合は、下記のメールでご連絡下さい。

******@*******.ac.jp

参加したくない場合は、署名してください。_________________

Figure N1. Japanese version of the informed consent form used in this study.
Note. Participants could decline to participate in the study or withdraw their data at any time. However, they still had to complete the course requirements, which included all the activities used for the study.

APPENDIX O

INTERVIEW INFORMED CONSENT FORM (ENGLISH)

Informed Consent

Title of Study: Japanese university students' learning experiences in English courses
Researcher: Aeric Wong, Ph. D. candidate in Applied Linguistics at Temple University Japan

Purpose of the Form

The purpose of this form is to ask you to participate in a research study. I will explain the study to you and ask you to volunteer to participate in the study. Please feel free to ask any questions that you may have before you decide if you want to participate. Whether you take part in the study is entirely up to you, and you are also free to withdraw from the study at any time. By signing this consent form, you are not waiving any legal rights that you have. You are only indicating that the study has been explained to you, that you have had an opportunity to ask questions, that you understand what is involved, and that you are willing to participate in the study.

Explanation of the study

The goal of this study is to investigate the language learning experiences of first-year university students and how they may affect their English and/or courses in their major area of study. This interview is part of my dissertation at Temple University Japan and will be published in the future. Data collected in the study will be used for educational and research purposes only.

Imposition on participants

The estimated duration of your participant in this study is 1 hour (two interview at approximately 30 minutes each. During this time, I would like to interview you about your language learning in various English courses. I would also like to record the audio of the interview. In total, participation in this study is estimated to involve 1 hour of your time.

Risks

The reasonably foreseeable risks or discomforts of participating in this study are minimal. However, if at any time, you do not feel comfortable answering a question, having your interview recorded, or having me observe you, please let me know and I will respect your requests.

Compensation:

To compensate you for your time, you will be paid 1,000 yen (minus income tax).

Your rights as a research participant

Your participation is voluntary. You may decide not to participate and you are free to stop participating at any time without penalty. If you decide not to participate at any time, no information about you will be included in this study.

Figure O1. English version of the informed consent form for interviewees, page 1.
Note. Participants did not see the English version. Space for questions by participants is omitted.

At any time now and in the future, you may ask me questions about this study and I will do my best to answer your questions to your satisfaction.

Confidentiality:

I will do my best to keep the sources of the data collected for this study confidential. I will keep all this information in a safe place and do my best to ensure your anonymity. All the proper names (names of people, institutions, places, etc.) used in this study, including yours, will be changed. I will do my best to limit the disclosure of your identity, but I cannot promise complete secrecy. There is always a potential risk of loss of confidentiality, and there is a slight chance that fellow researchers who hear your recorded voice might recognize you. In addition, in order to ensure that I am following the rules and regulations regarding research and the protection of human subjects, some organizations (e.g., the IRB at Temple University) may inspect and copy your information.

Your consent to participate in this study.

Please put a check mark next to the ways, if any, in which you agree to participate in this study.

1. I agree to participate in this research. ______
2. I agree to be audio recorded during the interview. ______
3. I also give permission to Aeric Wong to retain the audio-recordings for 12-months for educational and research purposes only. ______
4. I give permission to Aeric Wong to quote me using a pseudonym in future published papers. ______

I understand the information given to me, and I have received answers to any questions I have about the study. I understand and agree to the conditions of this study.

If you have any other conditions regarding the use of the data provided for this study, please specify them below.

I understand that my participation in this study is voluntary, and that I may withdraw from this study at any time by telling Aeric Wong to remove me from the study. Contact information is written below.

I also understand that I will receive a signed copy of this consent form.

____
Name (please print)

______________________________ ______________________________
Signature Date

Thank you very much for your participation.

Aeric Wong ******@*********.ac.jp

______________________________ ______________________________
Signature Date

Figure O1. English version of the informed consent form for interviewees, page 2.
Note. Participants did not see the English version. Space for questions by participants is omitted.

APPENDIX P

INTERVIEW INFORMED CONSENT FORM (JAPANESE)

説 明 同 意 書

研究タイトル: 日本の大学生の英語授業における学習体験について
研究者: エリック　ウォング。テンプル大学日本校、応用言語学博士候補生

本書の目的:
本書は、調査研究に参加して頂けるかをお尋ねするためのものです。調査者エリック ウォング（以下、「調査者」と記す）は、研究について説明をし、その研究へボランティアで参加して頂けるかを伺います。参加を希望する場合、参加を決める決める前に、不明点や疑問点について調査者にいつでも質問してください。参加するかどうかは、自由に決めてください。また、参加後いつでもこの調査から抜けることも可能です。この同意書にサインすることにより、あなたが法的権利を放棄するということではありません。あなたが研究について説明を受け、参加を決める前に質問をする機会があったこと、どんな影響を受けるかを理解し、そして本研究に自ら参加しているということを示すための同意書です。

研究について:
本研究のゴールは、大学 1 年生の言語学習体験を調査し、学生の英語コースおよび／または学部のコースにどのような影響があるかを調査することです。このインタビューは、テンプル大学日本校における調査者の論文の一部となり、論文は将来的に出版されることになります。本研究で集めたデータは、教育目的および研究目的にのみ使用されます。

参加者の負担:
本研究への参加にかかる予定所要時間は、1 時間（インタビュー2 回。各 30 分。）です。この時間を使って、様々な英語コースにおける、あなたの言語学習についてインタビューさせていただき、そのインタビューは音声録音をさせていただきます。2 回の合計で、約 1 時間のお時間を頂戴いたします。

リスク:
本研究に参加することによる、合理的に予知できるリスクや不快感はごく小さいものですが、もし質問に答えること、録音されていること、調査者があなたを観察していることに不快を感じることがあれば、いつでも申し出てください。あなたの希望を尊重します。

補償:
あなたの時間を使って参加して下さることに対しての埋め合わせとして、各インタビューの最後に 1,000 円（所得税別）をお支払い致します。

Figure P1. Japanese version of the informed consent form for interviewees, page 1.

参加者としての権利:
本研究への参加は任意(ボランティア)のものです。不参加を選ぶ場合もあると思います。参加したとしても、ペナルティーなしでいつでも自由に参加を取りやめることができます。もし参加しないと決めた場合、研究にはあなたに関する情報を入れません。
現在、またはこれから先、あなたが調査者に質問をすることがあると思いますが、その時は質問に対して最善の答えを提供し、満足いただけるように努めます。
守秘義務:
調査者は、本研究のために集めたデータソースを機密保持することに最善を尽くします。全情報を安全な場所で保管し、匿名性を確保することに努めます。あなたのものを含む、本研究で使用される本来の名前(氏名、学校、場所など)は、他のものに変更します。また、あなたを身元を開示することを制限することに最善を尽くしますが、すべての情報を完全に情報を機密保持することはお約束ができません。機密保持が失われるというリスクは常に存在し、インタビューで録音したあなたの声を同僚の調査者が聞き、あなたのことを特定できてしまうあるかもしれません。さらに、調査者が研究および対人実験の保護に関する規則および規定を守っていることを保証するために、組織団体(テンプル大学研究倫理審査委員会など)があなたの情報を調査し、コピーする場合もあります。

本研究への参加の同意:
下記の中で該当するもの全てにチェックマーク(レ)を入れてください。

1. 本研究への参加に同意します。 ____
2. インタビューが音声録音されることに同意します。 ____
3. 録音されたものが、教育・研究目的へのみ、エリック ウォング 12 か月間保管すること許可します。 ____
4. 将来発行される論文で、エリック ウォングが私の名前を偽名を使って引用する許可をします。 ____

私は、与えられた情報を理解し、研究に関する質問の答えを受け取りました。本研究の条件を理解し、同意します。

本研究のために提供されるデータの使用に関し、他に条件がある場合は、下記に明記してください。

Figure P1. Japanese version of the Informed consent form for interviewees, page 2.

私は、本研究への参加がボランティアであり、エリックウォングに伝えたうえで参加をいつでも辞めることを理解しました。連絡先は下記に記載。

この同意書のコピーを 1 枚受け取ることになっていることも理解しました。

____
名前（活字体で記入してください）

_______________________________　　　　_______________

署名　　　　　　　　　　　　　　　　　日付

ご参加ありがとうございます。

Aeric Wong

_______________________________　　　　_______________

Signature　　　　　　　　　　　　　　　Date

Figure P1. Japanese version of the Informed consent form for interviewees, page 3.

APPENDIX Q

TRACKING ESSAY SHEET

Tracking Essay

Student name: ____________________ Student number: ____________

Title: A memorable event in your life. **タイトル**：思い出に残る出来事

Directions: Write about a memorable event in your life. Describe who it was about and who you were with, what occurred, and when and where it took place. Write about it in the order it happened and give as much information as possible. Do not use a dictionary. You have 30 minutes. Please use the NGSL words you have studied in this writing.

説明: あなたのこれまでの人生において、思い出に残る出来事を１つ書いてください。誰についてか、あなたが誰と一緒だったか、何がいつ、どこで起きたかを説明してください。時間の流れに沿って、できるだけ多くの情報を入れて書いてください。辞書を使ってはいけません。時間は 30 分です。これまでに学んだ、NGSL ワードを使用してください。

Figure Q1. Reproduction of the Tracking Essay sheet, printed on an A3 piece of paper (equivalent to two 8 1/2 by 11 pages), used in this study.

APPENDIX R

NEW GENERAL SERVICE LIST TEST–PRODUCTIVE

Test of Written Productive Knowledge of the New General Service List	
There are 30 questions. You will read a Japanese word or phrase, and how it is used in a sentence. Write down the correct English word in the blank. Only the second letter is given. Only write one word. Be careful of the grammar.	
Example 1: 学校 これは大きい学校だ。 This is a big __c__________. Answer: This is a big sc**hool** .	***Example 2:*** 時間 彼らには時間がたくさんある。 They have a lot of __i__________. Answer: They have a lot of t**ime** .
1. 異なる；いろいろな それらは異なる。 They are __i__________.	7. 羽ばたいて飛ぶ 私たちは、そこへ飛行機で行きます。 We will ___l__________ there.
2. 上がる；起立する，上昇 それらは来週、上がります。 They will ___i__________ next week.	8. 解決策; 解答 これは、1 つの解決策だ。 This is one ___o__________.
3. 健康 健康は大事だ。 ___e__________ is important.	9. 圧力 ここに圧力をかけて。 Put ___r__________ on this.
4. 含む 私たちは、それを含んでいます。 We are ___n__________ it.	10. 利益、特典 これらが利益である。 These are the ___e__________.
5. 上手に、よく あなたは、それを上手にやった。 You did that ___e__________.	11. 試験をする；調べる 彼女は、これを調べたい。 She wants to ___x__________ this.
6. 考慮する；みなす 彼女はそれを考慮する。 She ___o__________ it.	12. 深い これは深い。 It is ___e__________.

Figure R1. Productive version of the New General Service List vocabulary test used in this study, page 1.

13. 余分の，余分なもの これらは余りの分です。 These are ___x__________.	22. 正当化する 私たちは、これを正当化することはできない。 We cannot ___u__________ this.
14. 範囲；広がり 彼女はそれの範囲を知らなかった。 She did not know the __x__________ of it.	23. 悲しい それは悲しい。 That is __a__________.
15. 保証，保証する 私はこれを保証します。 I ___u__________ this.	24. 影 影が見えます。 I can see the ___h__________.
16. 工場 これは新しい工場です。 This is a new ___a__________.	25. 動機 動機づけは重要だ。 ___o__________ is important.
17. 世界的な これは世界的な問題です。 This is a ___l__________ problem.	26. 徐々に 私はそれを徐々にやりました。 I did it ___r__________.
18. 枚；用紙 私は新しい用紙が必要です。 I need a new ___h__________.	27. 偏見 これは偏見です。 This is ___i__________.
19. 盗む これは盗まれたままです。 It has been ___t__________.	28. 多様性 私たちは、そこの多様性が好きです。 We like the ___i__________ there.
20. 課す；強いる 彼らは自分たちの考えを押しつけた。 They ___m__________ their ideas.	29. 喉 喉は大丈夫ですか？ Is your ___h__________ okay?
21. 曲線，曲がり これは曲線です。 This is a ___u__________.	30. 正確な これは正確ですか？ Is this ___c__________?

Figure R1. Productive version of the New General Service List vocabulary test used in this study, page 2.

APPENDIX S

VOCABULARY SELF-STUDY LOG

Name: _______________________________ Student Number: ___________

This is a self-study log for you to record the amount of time you spent studying vocabulary using (1) the Quizlet smartphone app, and/or (2) other methods such as checking the vocabulary sheet, practice writing the words, writing example sentences, using word cards, etc.

This will not affect your vocabulary course score so please be honest.

<u>Week 3</u>	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 4	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 5	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 6	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 7	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.

Figure S1. Vocabulary Self-Study log used to obtain general estimates of out-of-class time devoted to vocabulary study for this study, page 1.

Week 8	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 9	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 10	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 11	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 12	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 13	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 14	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.
Week 15	**Quizlet**	**Other**
How much time did you spend studying vocabulary yesterday?	_____ mins.	_____ mins.
How much time did you spend studying vocabulary last week?	_____ mins.	_____ mins.

Figure S1. Vocabulary Self-Study log, page 2

APPENDIX T

VOCABULARY INVENTORY SURVEY

NGSL 501-1000

Put a check mark next to the words you know.

Column 1	Column 2	Column 3	Column 4	Column 5
claim	stage	particularly	forget	guy
relationship	space	wife	page	function
travel	instead	sport	worry	yesterday
enjoy	realize	prepare	summer	image
death	usually	discuss	drink	size
nice	data	response	opinion	behavior
amount	single	voice	park	addition
improve	address	piece	represent	determine
picture	performance	finish	key	station
boy	chance	suppose	inside	population
regard	accept	apply	manager	fail
organization	society	president	international	environment
happy	technology	fire	contain	production
couple	mention	compare	notice	contract
act	choice	court	wonder	player
range	save	police	nature	comment
quality	common	store	structure	enter
project	culture	poor	section	occur
round	total	knowledge	myself	alone
opportunity	demand	laugh	exactly	significant
road	material	arm	plant	drug
accord	limit	heart	paint	wall
list	listen	source	worker	series
wish	due	employee	press	direct
therefore	wrong	manage	whatever	success
wear	foot	simply	necessary	tomorrow
fund	effort	bank	region	director
rest	attention	firm	growth	clearly
kid	upon	cell	evening	lack
industry	check	article	influence	review
education	complete	fast	respect	depend
measure	lie	attack	various	race
kill	pick	foreign	catch	recognize
serve	reduce	surprise	thus	window
likely	personal	feature	skill	purpose
certainly	ground	factor	attempt	department
national	animal	pretty	son	gain
itself	arrive	recently	simple	tree
teach	patient	affect	medium	college
field	current	drop	average	argue
security	century	recent	stock	board
air	evidence	relate	management	holiday
benefit	exist	official	character	mark
trade	similar	financial	bed	church
risk	fight	miss	hit	machine
news	leader	art	establish	achieve
standard	fine	campaign	indeed	item
vote	street	private	final	prove
percent	former	pause	economy	cent
focus	contact	everyone	fit	season

Figure T1. The Vocabulary Inventory Survey (as provided to participants), page 1.

Column 6	Column 7	Column 8	Column 9	Column 10
floor	husband	seat	fun	direction
stuff	access	throw	protect	facility
wide	movement	task	resource	variety
anyone	treat	normal	science	daily
method	identify	goal	disease	clothes
analysis	loss	associate	balance	screen
election	shall	blue	damage	track
military	modern	positive	basis	dance
hotel	pressure	option	author	completely
club	bus	box	basic	female
below	treatment	huge	encourage	responsibility
movie	yourself	message	hair	original
doctor	supply	instance	male	sister
discussion	village	style	operate	rock
sorry	worth	refer	reflect	dream
challenge	natural	cold	exercise	nor
nation	express	push	useful	university
nearly	indicate	quarter	restaurant	easily
statement	attend	assume	income	agency
link	brother	baby	property	dollar
despite	investment	successful	previous	garden
introduce	score	sing	dark	fix
advantage	organize	doubt	imagine	ahead
ready	trip	competition	okay	cross
marry	beyond	theory	earn	yeah
strike	sleep	propose	daughter	weight
mile	fish	reference	post	legal
seek	promise	argument	newspaper	proposal
ability	potential	adult	define	version
unit	energy	fly	conclusion	conversation
card	trouble	document	clock	somebody
hospital	relation	pattern	everybody	pound
quickly	touch	application	weekend	magazine
interview	file	hot	perform	shape
agreement	middle	obviously	professional	sea
release	bar	unclear	mine	immediately
tax	suffer	bill	debate	welcome
solution	strategy	search	memory	smile
capital	deep	separate	green	communication
popular	except	central	song	agent
specific	clean	career	object	traditional
beautiful	tend	anyway	maintain	replace
fear	advance	speech	credit	judge
aim	fill	dog	ring	herself
television	star	officer	discover	suddenly
serious	network	throughout	dead	generation
target	generally	oil	afternoon	estimate
degree	operation	dress	prefer	favorite
pull	match	profit	extend	difficulty
red	avoid	guess	possibility	purchase

Figure T1. Vocabulary Inventory Survey (as provided to participants), page 2

APPENDIX U

SEMI-STRUCTURED INTERVIEW QUESTIONS

Pre-interview explanation

Purpose is to collect data, not to assess Ss performance (i.e., give grades or points). Responses will not affect course grades and responses will not be shared with course instructors. So please be honest.

For FIRST interview (July, 2017)

Could you tell me about vocabulary course 1?

1) What did you like best about vocabulary course 1?
2) What did you find most difficult about this class?
3) How did your vocabulary study change during this class?
4) How do you study vocabulary now? How did you study it in high school? If different, why did you change to the style now?
5) What do you try to remember when you study vocabulary? (e.g., L1 or L2 translation, meaning, spelling, word in a sentence)
6) When did you study vocabulary? (e.g., after class, in the evenings, on weekends)
7) Last week, how much time did you spend on vocabulary study? (time estimate research shows that this is generally accurate).

8) How much time did you spend on it every week? (time estimate research shows this will match the time they spent in the most recent week).

9) I want to show you three phrases. Please select the one that best describes when you study vocabulary.

    a. I ONLY studied vocabulary just before the vocabulary tests.

    b. I studied vocabulary ONE or TWO times a week.

    c. I studied vocabulary at least ONE time EACH day.

10) Show Self-study Log - You were asked to record how much time you studied vocabulary every week. How much do you think your study time affected your vocabulary learning?

11) Looking at your vocabulary surveys, how many new words did you learn over this semester? (How has your vocabulary changed during this semester?)

12) You studied NGSL Column 1, 2, and 3 in the first few weeks of the semester and studied the words you didn't know. (Show Pre- and/or Post-Vocabulary Inventory). How well can remember those (previously unknown) words now?

Tell me about content course 1.

1) What were the easiest things for you in content course 1?

2) What were the most difficult things?

3) What did your teacher do in this class? (Describe a typical lesson)

4) The vocabulary course 1 teachers asked you and the other students to study the NGSL words because they will help you in all your English classes. How much did the vocabulary study in this course help you in your Procedure, Report, and Narrative writing?

5) How frequently did you try to use the NGSL words in your Procedure, Report, and Narrative?

Wrap-up

1) Please explain how the style of classes here are different than your high school.

2) What did you like about the English classes here?

3) What didn't you like about the English classes here?

4) Tell me about how your overall English level has changed.

5) How do you think you will use English in the future?

For LAST interview (January, 2018)

1) How has your English vocabulary changed during this semester?

2) [All conditions]: We (the teachers) have asked you and the other students to study these words because they will help you in all your English classes.

   a. Please tell me how you study vocabulary now. How did you study it in high school?
   b. Why do you use this style of studying vocabulary now?
   c. What information do you try to remember when you study vocabulary? (e.g., L1 or L2 translation, meaning, spelling, word in a sentence)
   d. How much time do you spend studying vocabulary?
   e. When did you study vocabulary?
   f. I want to read you three phrases. Please select the one that best describes when you study vocabulary.
      i. I ONLY study vocabulary just before the vocabulary tests.
      ii. I studied vocabulary ONE or TWO times a week.
      iii. I studied vocabulary at least ONE time EACH day.
   g. Please tell me how useful you felt the words you studied were in the writing you did for the Content Course.
   h. In your writing, how frequently did you try to use the words you were studying?

3) [Tracking Essay Comparison]

   a. I am going to show you your tracking essays from April and July. Before I do, please tell me how you think your writing has changed.
   b. Looking at your tracking essays (April-July), how would you evaluate the changes in your writing?
   c. Do you think your writing has improved? Why or why not?

4) [R, R+, and R++ conditions]

   In your Content Course, your teacher reminded you to use the vocabulary in your homework and your essays.

   a. What effect did the reminders from the Content Course teacher have on your vocabulary study?
   b. Did the reminders help you use more vocabulary in the Content Course?
   c. What effect did the reminders from the Content Course teacher have on your use of the words in your essays?

5) [R+ Condition]

   a. In your Content Course, you had to look at the vocabulary lists and circle the words that could be used in your essay. How useful was this in helping you to use the vocabulary you studied?

b. Did you try to use words that you studied before but were not in the vocabulary list the teacher gave you? (e.g., given only Column 7, did they use words from Columns 1-6?)

6) [R++ Condition]

   a. In your Content Course, you had to look at the vocabulary lists, circle the words that could be used in your essay, write them in the box at the top of the paper for the final draft, then revise your essay using these words. How useful was this in helping you use the vocabulary you studied?

7) [Tracking Essay Comparison]

   a. I am going to show you your tracking essays from April, July, and January. Before I do, please tell me how you think your writing has changed.

   b. Looking at your tracking essays (April-July-January), how would you evaluate the changes in your writing?

   c. How much do you think you used the vocabulary you studied this term in your tracking essays?

APPENDIX V

TECHNIQUE FEATURE ANALYSIS SCORES OF INSTRUMENTS

Table V1. *Technique Feature Analysis Score - Tracking Essays*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	0
	Does the activity motivate learning?	0
	Do the learners select the words?	1
Noticing	Does the activity focus attention on the target words?	0
	Does the activity raise awareness of new vocabulary learning?	0
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	1
	Is it productive retrieval?	1
	Is it recall?	1
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	1
	Is it productive?	1
	Is there a marked change that involves the use of other words?	1
Retention	Does the activity ensure successful linking of form and meaning?	0
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		7

Table V2. *Technique Feature Analysis Score - Receptive New General Service List Test (NGSLT-Receptive)*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	0
Noticing	Does the activity focus attention on the target words?	1
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	0
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	0
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	1
Total Score		4

Note. NGSLT-Receptive = New General Service List Test-Receptive.

Table V3. *Technique Feature Analysis Score - Productive New General Service List Test (NGSLT-Productive)*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	0
Noticing	Does the activity focus attention on the target words?	0
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	1
	Is it productive retrieval?	1
	Is it recall?	1
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	0
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		5

Note. NGSLT-Receptive = New General Service List Test-Productive.

Table V4. *Technique Feature Analysis Score - Quizlet Self-study*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	1
	Do the learners select the words?	0
Noticing	Does the activity focus attention on the target words?	1
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	1
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	1
	Is there spacing between retrievals?	1
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	0
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	1
Total Score		8

Table V5. *Technique Feature Analysis Score - Moodle Self-study*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	0
Noticing	Does the activity focus attention on the target words?	1
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	1
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	0
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	1
Total Score		5

Table V6. *Technique Feature Analysis Score - Writing Assignments – No Reminder Group (NR)*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	0
	Does the activity motivate learning?	0
	Do the learners select the words?	1
Noticing	Does the activity focus attention on the target words?	0
	Does the activity raise awareness of new vocabulary learning?	0
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	0
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	1
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		2

Note. NR = No Reminder group.

Table V7. *Technique Feature Analysis Score - Writing Assignments – Reminder Group (R)*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	1
Noticing	Does the activity focus attention on the target words?	0
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	0
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	1
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		4

Note. R = Reminder group.

Table V8. *Technique Feature Analysis Score - Writing Assignments – Remind + Notice (R+) Group*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	1
Noticing	Does the activity focus attention on the target words?	1
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	0
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	0
	Is it productive?	0
	Is there a marked change that involves the use of other words?	0
Retention	Does the activity ensure successful linking of form and meaning?	1
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		5

Note. R+ = Remind + notice group, R++ = Remind + use group.

Table V9. *Technique Feature Analysis Score - Writing Assignments – Remind + use (R++) Group*

	Criteria	Scores
Motivation	Is there a clear vocabulary learning goal?	1
	Does the activity motivate learning?	0
	Do the learners select the words?	1
Noticing	Does the activity focus attention on the target words?	1
	Does the activity raise awareness of new vocabulary learning?	1
	Does the activity involve negotiation?	0
Retrieval	Does the activity involve retrieval of the word?	0
	Is it productive retrieval?	0
	Is it recall?	0
	Are there multiple retrievals of each word?	0
	Is there spacing between retrievals?	0
Generation	Does the activity involve generative use?	1
	Is it productive?	1
	Is there a marked change that involves the use of other words?	1
Retention	Does the activity ensure successful linking of form and meaning?	1
	Does the activity involve instantiation?	0
	Does the activity involve imaging?	0
	Does the activity avoid interference?	0
Total Score		8

Note. R++ = Remind + use group.

APPENDIX W

REMINDER POSTER

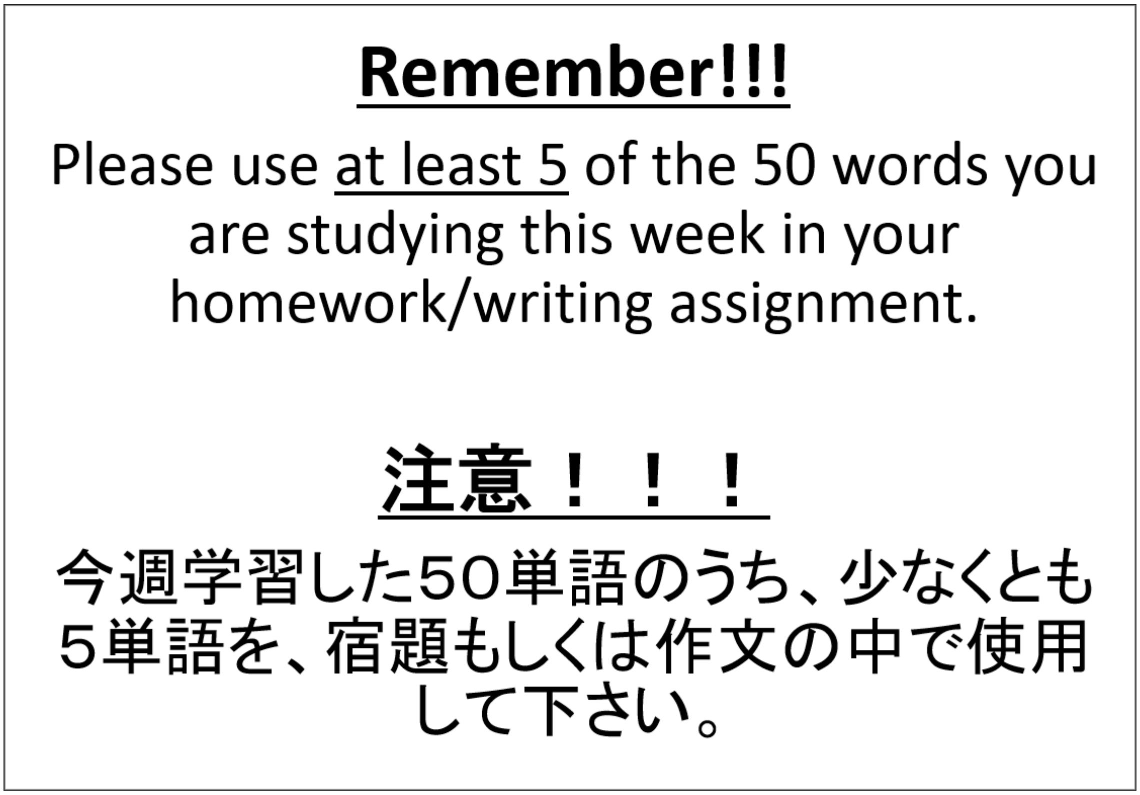

Figure W1. Reproduction of the reminder poster, printed on an A3 piece of paper (equivalent to two 8 1/2 by 11 pages), that was shown to the three treatment group R, R+, and R++ participants during this study.

Note. *R = Reminder group, R+ = Remind + Notice group, R++ = Remind + Use group*

APPENDIX X

VOCABULARY LIST RECEIVED BY THE R+ PARTICIPANTS

NGSL 501-1000
Column 4

forget	necessary
page	region
worry	growth
summer	evening
drink	influence
opinion	respect
park	various
represent	catch
key	thus
inside	skill
manager	attempt
international	son
contain	simple
notice	medium
wonder	average
nature	stock
structure	management
section	character
myself	bed
exactly	hit
plant	establish
paint	indeed
worker	final
press	economy
whatever	fit

Figure X1. Reproduction of the vocabulary list given to the R+ participants in the content course of the words studied during the previous week in the vocabulary course during this study.

APPENDIX Y

MATERIALS RECEIVED BY THE R++ PARTICIPANTS

NGSL 501-1000

Column 1	Column 2	Column 3	Column 4
claim	stage	particularly	forget
relationship	space	wife	page
travel	instead	sport	worry
enjoy	realize	prepare	summer
death	usually	discuss	drink
nice	data	response	opinion
amount	single	voice	park
improve	address	piece	represent
picture	performance	finish	key
boy	chance	suppose	inside
regard	accept	apply	manager
organization	society	president	international
happy	technology	fire	contain
couple	mention	compare	notice
act	choice	court	wonder
range	save	police	nature
quality	common	store	structure
project	culture	poor	section
round	total	knowledge	myself
opportunity	demand	laugh	exactly
road	material	arm	plant
accord	limit	heart	paint
list	listen	source	worker
wish	due	employee	press
therefore	wrong	manage	whatever
wear	foot	simply	necessary
fund	effort	bank	region
rest	attention	firm	growth
kid	upon	cell	evening
industry	check	article	influence
education	complete	fast	respect
measure	lie	attack	various
kill	pick	foreign	catch
serve	reduce	surprise	thus
likely	personal	feature	skill
certainly	ground	factor	attempt
national	animal	pretty	son
itself	arrive	recently	simple
teach	patient	affect	medium
field	current	drop	average
security	century	recent	stock
air	evidence	relate	management
benefit	exist	official	character
trade	similar	financial	bed
risk	fight	miss	hit
news	leader	art	establish
standard	fine	campaign	indeed
vote	street	private	final
percent	former	pause	economy
focus	contact	everyone	fit

Figure Y1. The vocabulary list provided to the R++ participants in the content course.
Note. R++ = Remind + use group.

Procedure Genre Writing
Name: ______________________________   Topic: How to ______________________________
Useful vocabulary from NGSL Columns 1-4
Final Draft   When you write the final draft of your procedure, please check the comments from your peer review evaluation and use the NGSL words you wrote in the box above.

Figure Y2. Reproduction of the worksheet given to the R++ participants for them to write down words and revise their essays during this study.
Note. R++ = Remind + use group.