EFFECTS OF WORD CARD METHODOLOGY AND TESTING ON VOCABULARY KNOWLEDGE AND MOTIVATION

A Dissertation Proposal
Submitted
to the Temple University Graduate Board

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

By
Darrell Wilkinson
August 2020

Examining Committee Members

David Beglar, Advisory Chair, Teaching and Learning
Tomoko Nemoto, Teaching and Learning
Paul Nation, External Member, Victoria University of Wellington
Batia Laufer, External Member, University of Haifa

ABSTRACT

This research explores the efficacy of word cards as a deliberate vocabulary learning technique and investigates the effects of testing on motivation for deliberate vocabulary study. While word cards are a widely supported method of deliberately studying target foreign language vocabulary, there is a surprising lack of research-based evidence supporting them, and very few researchers have asked students how they feel about word card learning. In addition, it is well known that testing can increase extrinsic motivation, but the motivational effects of regular vocabulary testing has so far been largely overlooked.

The first two experiments in this study investigated the learning outcomes associated with making and studying from word cards. Experiment 1 is unique in that it is the first known study to isolate the learning outcomes of making word cards. The results indicated that the process of making word cards results in significant initial vocabulary learning, but a substantial amount of this new knowledge is lost in a relatively short period of time if no further study is carried out. The second experiment compared the use of self-made word cards with premade cards. The results indicated that although both methods are effective in the short and long-term, learners might be better studying from premade cards. However, the results of the third experiment led to the conclusion that premade cards are best suited to motivated learners. Qualitative data analysis revealed that learners generally have positive views of word card learning and understand the benefits it offers over other methods such as learning from lists or notebooks.

The final experiment in this study examined the motivational effects of regular student-generated vocabulary testing and investigated whether learners can act as reliable raters of their peers' vocabulary development. Results indicated that peer-to-peer vocabulary testing increases motivation for deliberate vocabulary learning even if test scores do not affect the students' grades. Surprisingly, the student-generated and administered tests were more intrinsically motivating than extrinsically so.

Taken together, the results of the various experiments offer support for the use of word cards for foreign language vocabulary learning and indicate that motivation for such types of learning can be increased by through peer testing.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. David Beglar for the continuous and invaluable support throughout my Ph.D. study and related research. His patience, broad knowledge, attention to detail, and encouragement were invaluable.

Besides my advisor, I would like to thank the rest of my examining committee: Prof. Paul Nation, Dr. Batia Laufer, and Dr. Tomoko Nemoto for their insightful comments and encouragement.

Last but not least, I would like to thank my partner, Natalie Close, for supporting me throughout my Ph.D. studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Importance of Vocabulary to Language Learning and Use

There is no doubt that vocabulary knowledge plays a vital role in learners' ability to proficiently use a foreign language (Carter & McCarthy, 2014, pp. 201–203; Folse, 2011; Laufer & Shmueli, 1997; Nation, 2013, pp. 14–16) As Schmitt (2008) stated, "One thing that students, teachers, materials writers, and researchers can all agree upon is that learning vocabulary is an essential part of mastering a second language" (p. 329). A more definitive statement comes from Wilkins (1972) who stated, "While without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (p. 111).

Numerous researchers have indicated the importance of vocabulary to successful reading comprehension (Anderson & Freebody, 1981, pp. 77–171; Laufer, 1992a, 1992b, pp. 126–132; Nation, 2006; Qian, 2002), writing ability (Astika, 1993; Engber, 1995; Laufer & Nation, 1995; Olinghouse & Leaird, 2009), and speaking ability (Webb, 2009). Therefore, the domain of vocabulary acquisition is central to the wider arena of second language acquisition (SLA), and as such, second or foreign language learners need to devote considerable amounts of time and effort to learning target language vocabulary (Folse, 2006; Nation, 2013, pp. 216–218, 451–453; Schmitt, 2000, p. 4).

Main Vocabulary Learning Approaches: Incidental and Intentional

Vocabulary learning activities are usually designed to allow vocabulary to be learned either incidentally or intentionally. The current widely adopted view is that both incidental and intentional vocabulary learning approaches are necessary if learners are to acquire the many thousands of words needed for communication in a foreign language (Folse, 2011, Laufer, 2003; Nation, 2013, pp. 2–3; Schmitt, 2008).

Incidental vocabulary learning is also referred to as implicit or indirect vocabulary learning. Incidental vocabulary learning involves learners acquiring vocabulary knowledge in a natural context, for example, while engaged in other activities such as reading (Ellis, 1995). A further definition is provided by Hulstijn (2001) who described incidental vocabulary learning as “the learning of vocabulary as the by-product of any activity not explicitly geared to vocabulary learning” (p. 271). Rieder (2003) stated that incidental vocabulary acquisition can be regarded as non-explicit because it does not involve an explicit vocabulary learning intention; the learner’s goal is text comprehension, not vocabulary acquisition. In other words, as Han and Cheng-ling (2010) pointed out, vocabulary acquired incidentally is a result of the need for text comprehension rather than a focus in its own right.

Vocabulary knowledge acquired incidentally often involves more than just form and meaning connections, but also raises knowledge of underlying lexical features and colocations (Ellis, 2004). According to Ellis (1995), the learner often might not be aware that they have acquired this new vocabulary knowledge, thus, the vocabulary knowledge gained is described as implicit knowledge. Implicit vocabulary acquisition underlies

Krashen’s Input Hypothesis (1989), as can be seen from Krashen’s statement that “language is subconsciously acquired—while you are acquiring, you don’t know you are acquiring; your conscious focus is on the message, not form” (p. 440). Also, according to Ellis (1995) and Krashen (1989) the acquired vocabulary knowledge continues to be subconsciously held in the brain even though the learner might have no conscious knowledge concerning vocabulary items they have acquired implicitly.

However, Ellis (1994) claimed that both implicit and deliberate learning mechanisms are often involved in incidental vocabulary acquisition (pp. 1–31). This means that while, as described above, learners mainly learn unconsciously through repeated meeting in various contexts, they might sometimes temporarily switch to a deliberate focus, for example, looking up an unknown word found in a reading passage. Nevertheless, the main focus remains on gaining comprehension of the text rather than explicitly studying vocabulary.

Research on incidental vocabulary learning has shown that words are gradually learned through repeated encounters in context, and the more often unknown words are encountered, the more likely they are to be learned (Elley, 1991; Ellis, 2002; Laufer, 2003; Waring & Takaki, 2003). Heidari-Shahreza and Tavakoli (2012) pointed out that it is becoming widely accepted that more explicit approaches to vocabulary learning and teaching can complement and reinforce implicit vocabulary acquisition.

Intentional vocabulary learning is often also termed as deliberate or explicit vocabulary learning and sometimes as language-focused learning (Elgort, 2011; Schmitt, 2008; Nation, 2103, p. 216). Intentional vocabulary study involves learners carrying out activities with the sole purpose of acquiring vocabulary and includes activities such as studying from word lists or word cards, and carrying out vocabulary focused reading, writing, speaking, and listening activities (Elgort, 2011; Schmitt, 2008; Webb, 2020, p. 5). For example, if learners are writing a paragraph or essay with the purpose of using specified target words, this would still be classified as intentional vocabulary activity. Hulstijn (2001), described intentional vocabulary learning as “any activity aiming at committing lexical information to memory” (p. 271). According to Ellis (2015), during activities such as deliberate vocabulary learning activities, knowledge is gained through a conscious process refereed to a selective learning; where the individual attends to particular aspects of the lexical items and tests hypotheses in a search for underlying structures, regularities and rules (p. 3).

Deliberate vocabulary learning activities are believed to produce explicit vocabulary knowledge; lexical knowledge learners are conscious of (Ellis, 2015, Hulstijn, 2011). Explicit acquisition shows itself in that learners can verbalize and explain their reasoning behind the assumptions they have made about lexical meaning or rules. Explicit acquisition of vocabulary can result from any type of activity, as long as the learner consciously stops and makes an effort to understand a particular aspect of lexical items encountered during or after the activity. According to Ellis (1995), learners’ explicit acquisition of new vocabulary can be strongly facilitated by the use of a range of metacognitive strategies such as noticing, inference, and consolidating knowledge through repeated study.

The Need for Both Types of Learning

Just as there is no one best approach to facilitate successful language learning, there is no one best approach to teaching and learning vocabulary. While incidental vocabulary activities are beneficial for consolidating previous knowledge and learning more information about previously encountered words, deliberate vocabulary study is a quick and efficient method of learning vocabulary (Heidari-Shahreza & Tavakoli, 2012; Nation, 2013, p. 217; Schmitt, 2008, Webb, 2020, pp. 5–6). All learners acquire vocabulary incidentally and most will also choose to do so deliberately, and these different types of approaches and knowledge interact and play their own part in future language learning, comprehension, and use (Ellis, 2015, p. 4). As Heidari-Shahreza and Tavakoli (2012) pointed out, it is becoming widely accepted that more explicit approaches to vocabulary learning and teaching can complement and reinforce implicit vocabulary acquisition, or vice versa.

Table 1. *Summary of Key Terminology*

Term	Definition
Incidental learning	Learning new information unconsciously.
Intentional learning	Making a specific and conscious effort to learn something.
Incidental vocabulary learning	Learning vocabulary as a by-product of other activities such as reading or listening.
Intentional/Deliberate vocabulary learning	Engaging in activities with the sole purpose of acquiring vocabulary

Statement of the Problem

Overall, while there is much research on deliberate learning techniques, there is a surprising lack of research comparing the efficacy of different word card designs. Many studies have utilized word cards while investigating other issues affecting vocabulary acquisition (Elgort, 2011; Mondria & Wiersma, 2004, pp. 79–100; Nakata, 2015; Waring, 1997). Also, some studies have compared the efficacy of word cards with other deliberate study materials such as lists (Kitano & Chiba, 2019; Kuo & Ho, 2012; Nakata, 2008), with vocabulary instruction (Komachali & Khodareza, 2012; Reynolds & Shih, 2019), and digital applications (Emrah & Omur, 2010; Nakata, 2008). However, very few studies were found that compared different word card approaches, and those did either investigated different digital word card designs (Dodigovic, 2013; Nakata, 2010), or compared digital word cards with paper cards (Nikoopour, & Kazemi, 2014; Oberg, 2011). In addition, some of the studies (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Oberg, 2011) have serious methodological or data reporting flaws that make it difficult to rely on the findings.

Therefore, the first specific problem due to the above is that although there is considerable support for the use of word cards, no researchers have shown how much learning occurs from the action of simply making word cards, or how sensitive this knowledge is to attrition if no further study is carried out (Nation & Webb, 2011, p. 42). Making these types of study materials can take a considerable amount of time and effort on the part of the learner, but it is unknown if this process is necessary. Considering the time-consuming nature of making word cards, and given the fact that while learners need to learn thousands of words, there are many aspects of language that learners need to focus on other than vocabulary so issues such as time-on-task and task efficiency are important (Folse, 2011; Laufer, 2014; Pichette, de Serres, & Lafontaine, 2011; Webb, 2002, 2005). Nation and Webb (2011) stated, "We know from experience that after making the cards, learners can recall from 30-70% of the meanings on the first run through the cards" (p. 42), however, empirical research supporting this assertion has not been found. No studies have been found that isolated the vocabulary learning outcomes associated with making word cards or showed how robust this knowledge is to attrition.

The second specific problem is related to the above and is that while we can presume that making word cards will result in positive learning outcomes, little research has been carried out examining if it is more beneficial to have students make their own word cards or be given pre-made cards. Only one such study was found (Dodigovic, 2013), but this study compared premade and self-made digital word cards and suffers from some serious research design flaws. In most of the word card studies carried out to date, researchers have compared the efficacy of word cards with other study methods (Kitano & Chiba, 2019; Kuo & Ho, 2012; Nakata, 2008; Reynolds & Shih, 2019), or have focused on digital word cards versus traditional paper-based cards (Emrah & Omur, 2010; Nikoopour, & Kazemi, 2014; Oberg, 2011). No research was found that compared the learning outcomes of making traditional word cards with that of studying from pre-made cards for the same amount of time, therefore, we do not know whether the substantial amount of time needed to make word cards is time well spent or not.

Moving away from word cards, another major problem is the lack of empirically supported techniques available to help researchers and teachers overcome the difficulty of testing learners on their self-selected target vocabulary items. The connection between the vocabulary self-selection strategy and vocabulary testing has not been examined in the literature to date. Having students select the individual lexical items they will study, usually from reading passages or frequency-based lists, is a widely recommended method of vocabulary selection (Brown, 2012; Cobb, 1999; Mizumoto, 2013; Reynolds & Shih, 2019). However, while self-selection might well be a pedagogically sound vocabulary learning approach, it opens up specific challenges regarding vocabulary testing that have not been dealt with in the literature. For example, if learners in the same class self-select different words from the same list, or study from different lists based on their goals and needs (Hirai, 2014), then how can they be assessed? Assessment becomes extremely difficult because if students are self-selecting their target vocabulary, they will each have a unique set of vocabulary study items, therefore, administering standardized vocabulary tests is unproductive, and currently there is not clear way to regularly test all students on their individual target vocabulary knowledge. However, no research was found that introduced a simple and time effective means of testing individual learners' vocabulary development repeatedly during a course. Such a tests system could be a valuable tool for both classroom vocabulary assessment, especially if it can be found to be reliable, and valuable for providing students feedback on their deliberate vocabulary study progress. Therefore, the fact that no such system is available is a major problem.

In addition, while the positive relationship between testing and extrinsic motivation has been examined and discussed (Butler, 2010; Harackiewicz & Sansone, 2000, pp. 79–104; Karpicke, & Roediger, 2008; Roediger, & Karpicke, 2006a, 2006b), the lack of research examining the effects of regular testing on motivation for deliberate vocabulary learning represents another problem. As, Tseng and Schmitt (2008) stated, "only a small number of studies have been undertaken to examine the role of motivation in vocabulary learning" (p. 358), and this is especially true in relation to vocabulary testing and motivation. No empirical L2 vocabulary research has compared the effects of different test conditions on aspects such as motivation and time on task. In addition, qualitative data assessing students' perceptions of regular vocabulary testing, and how these perceptions affect their motivation or time-on-task has not been found. As Schmitt (2008) stated, long-term and continued motivation is vital for vocabulary development, but little is known about how this can be achieved. This represents a significant problem as learners often have low levels of intrinsic motivation for deliberate vocabulary learning activities (Al-Murtadha, 2014; Ng & Ng, 2015; Wilkinson, 2017).

An additional significant problem concerns the instruments and measures that many previous vocabulary studies have used to assess (a) vocabulary knowledge and (b) motivation. First, as noted by Nation and Webb (2011, pp. 102–104) and Webb (2009), much of the research on both incidental and intentional vocabulary study has examined learning effects purely in terms of one form of recall, either passive or active, but often not both. In addition, very few deliberate vocabulary learning studies award partial credit, with answers simply being marked as correct or incorrect, thus ignoring any possible partial learning (Barcroft, 2009). Therefore, the narrow approach of assessing vocabulary knowledge gains used in many L2 vocabulary studies is a serious problem because, as noted by Webb (2009) and Laufer and Rozovski-Roitblat (2014), valuable information regarding the varying degrees of knowledge gained from a particular treatment is not represented in the results. For example, while learners might not be able to fully or correctly recall and write a target vocabulary item, they might well have learned something about the word form or meaning that the tests were not able to pick up (Webb, 2009). In addition, much of the literature on motivation has relied on self-report data rather than utilizing concrete measures of motivation such as time-on-task. Relying on self-report data from sources such as surveys and interviews alone is not recommended because such data are often unreliable and should be triangulated with other data (Creswell, 2014, pp. 199–210; Hatch, 2002, pp. 93–133; Mackey & Gass, 2005, p. 78).

Another major problem concerning current vocabulary research methodology is that few L2 vocabulary researchers have combined carefully controlled cross-sectional studies with more longitudinal studies (Nakata, 2020, p. 314). Cross-sectional research can control for confounding variables very well and thus isolate treatment effects, but a major problem is that they often represent unnatural conditions and unlike like longitudinal designs, they do not offer insights into what happens in real-life contexts over time (Creswell, 2014; pp. 12–15; Mackey & Gass, 2005, p. 204). On the other hand, while longitudinal data better represent what happens in actual natural learning contexts, it is very difficult to control or assess the impact of confounding variables not under investigation (Creswell, 2014, p. 12; Mackey & Gass, 2005, pp. 111, 161, 303).

Therefore, while more longitudinal word card studies are needed (Nakata, 2020, p. 314), vocabulary research designs that combine the two approaches offer the benefits of both. Thus, the lack of vocabulary research designs that combine both cross-sectional and longitudinal vocabulary treatments represents a problematic gap in the vocabulary learning literature.

The final problem is that few researchers investigating deliberate vocabulary study methods such as word cards have employed qualitative or mixed methods approaches (Nation & Webb, 2011, p. 43; Webb, 2020, p. 7). Some vocabulary studies have reported the use of qualitative data to support or triangulate the quantitative data, but most that have done so have largely relied on one method of qualitative data collection, usually questionnaires. Few vocabulary researchers have followed the advice by experts in the field (Creswell, 2014, p. 200; Creswell & Clark, 2017, pp. 54–55; Hatch, 2002, pp. 93–133; Mackey & Gass, 2005, pp. 201–202) and triangulated data by utilizing a variety of qualitative methods such as interviews, observations, questionnaires, and artifact checking. This is a serious problem as the insights that can be garnered from such qualitative data can highlight various factors that might have affected the learning outcomes, for example, individual learner differences and teacher factors (Nation & Webb, 2011, p. 43; Webb, 2020, p. 7).

Purposes and Significance of the Study

The overall purpose of this study is to address various gaps in the literature regarding the efficacy of using word cards for deliberate vocabulary learning, and to explore issues relating testing and motivation for deliberate study. This study not only provides a thorough and up-to-date review of the literature on deliberate L2 vocabulary learning, especially with regard to word cards, but also adds to the literature by providing further empirical results regarding deliberate L2 vocabulary study using word cards. In addition, the third experiment in this study offers new insights into the effects of various test conditions on motivation. Therefore, the results of the study will be of interest and importance to researchers and educators interested in L2 vocabulary acquisition in general, and more specifically to those concerned with deliberate L2 vocabulary study, motivation, and assessment.

The first purpose of this study is to investigate how much learning takes place from simply making word cards and establish how sensitive this knowledge is to attrition. Although Nation and Webb (2011) have speculated how much learning occurs from making such materials (p. 42), no empirical evidence supports these suppositions. This research is designed to address this gap by empirically showing how much learning takes place from making word cards. More specifically, I aim to add to the literature by examining how much knowledge of target vocabulary learners recall productively and receptively after making word cards, and demonstrate how sensitive this knowledge is to attrition over time. As there is no current empirical research regarding the amount of vocabulary that is learned and retained as a result of making word cards, the results fill a specific and important gap in the L2 vocabulary literature.

The second purpose of this research is to cross-sectionally and longitudinally compare the efficacy of making and studying from word cards with simply studying from pre-made versions for the same amount of time. Whether or not the considerable amount of time learners spend on making word cards is of more benefit than spending the same amount of time simply studying ready-made versions is an important question that has so far not been widely addressed in the literature. Therefore, this study is the first to compare the learning outcomes from making and studying word cards with that of studying from pre-made equivalents for the same amount of time. I hope to indicate whether the considerable amount of time needed to make materials is time well spent for learners, or whether it is simply an accepted and unchallenged pedagogical practice based more on convenience than effectiveness.

The third aim of this study is to offer evidence in favor of a peer-to-peer test system that could prove beneficial in contexts where learners are self-selecting their target vocabulary items. A major problem facing teachers is that if the self-selection of target items is to be encouraged in language classrooms, but regular testing is also viewed as being important, then teachers are left with the challenge of how to quickly and regularly test learners on their individual vocabulary sets. To the best of my knowledge, no empirically supported solution is currently available, therefore, this study aims to offer a practical solution to this challenge. If the tests used in this study can be shown to be a reliable and useful classroom assessment tool that allows learners to be quickly, efficiently, and regularly tested on their individual vocabulary sets, this could have significant pedagogical implications. Vocabulary self-selection is recommended by L2 vocabulary experts (Nation, 2013, pp. 98–104; Schmitt, 2008) and is supported by research (Cobb, 1999; Masoudi, 2017; Moir & Nation, 2002; Ogawa 2012). In addition, the benefits of repeated testing are well known (Butler, 2010; Roediger, & Karpicke, 2006a, 2006b), therefore, this study aims to offer a validated system for testing students on self-selected target vocabulary.

Furthermore, a fourth purpose of this study is to examine the effects of different test conditions on motivation. As noted above, regular testing is believed to increase both motivation and learning outcomes (Butler, 2010; Deci, 1971; Harackiewicz & Sansone, 2000, pp. 79–104; Roediger, & Karpicke, 2006a, 2006b), but this study uses quantitative and qualitative data to assess the effects of various test conditions on motivation for deliberate vocabulary learning. The study is significant because it is the first known study to show the motivational effects resulting from the inclusion of peer-generated vocabulary testing into language curricula.

Another aim of this study is to examine all the issues above using more sensitive and accurate measures of vocabulary knowledge and motivation than used in many previous studies. As mentioned, much of the research on word cards have examined learning effects using only one test or using only a correct/incorrect measuring system. Therefore, it is hoped the vocabulary measures used in this study more accurately represent the learning outcomes attributable to the treatments. In addition, by using more accurate measures of motivation than self-report data, this research aims to provide more concrete motivational outcomes attributable to regular vocabulary testing.

In addition, a further aim of this study is to provide not only carefully controlled cross-sectional data, but to also compare such results with longitudinal data generated from more authentic and less controlled treatments. The careful research design and cross-sectional data isolate the short-term effects of the word card treatments, but the longitudinal data provide more ecologically valid data regarding the longer-term results of the various methods in more natural learning contexts.

Finally, by utilizing a mixed-methods design that includes not only statistical analysis but analysis of a variety of qualitative data gained from observations, interviews, surveys, and artifact checking, I aim to offer insights into the various individual and group factors that might have affected the outcomes. The additional qualitative data support and add to the knowledge gained from the quantitative data, providing much more detailed information into not only what happened, but also how and why.

In conclusion, overall, this research is designed to answer a number of specific and important questions that have yet to be addressed in the second language vocabulary acquisition literature. This research first offers empirical evidence showing the learning outcomes of simply making word cards. In addition, by comparing the learning outcomes attributable to student-created word cards with that of studying from teacher-made equivalents for the same amount of time, new insights are provided about whether or not current widely adopted practices are pedagogically sound. Furthermore, the analysis of the student-to-student test system aims to offer data that enable teachers to decide if they can use the test system as a means of assessing learners' individual vocabulary development, or for increasing motivation.

The Audience for the Study

A number of groups of people will benefit from this study including researchers, second language teachers, and curriculum/materials designers. First, as the research addresses specific gaps in the literature regarding developing L2 vocabulary knowledge, and contains enough information regarding research design and analysis to facilitate replication of the study, it might be of benefit and interest to not only other researchers concerned with L2 vocabulary acquisition, but also to those involved with other areas of second language acquisition and pedagogy research.

In addition, as the deliberate vocabulary study approaches and assessment methods can be easily included in wide range of language courses, it is also of clear benefit and interest to language course and curriculum designers, materials designers, and language teachers. The results will allow the above groups of people to make more informed choices about how learners can best achieve their lexical acquisition goals over time.

Finally, as cross-sectional and longitudinal data are collected and the research combines both quantitative and qualitative techniques, the study will be of interest to researchers with an interest in such cross-sequential mixed-methods research designs.

Delimitations

This study has a number of delimitations in terms of objectives, participants, and learning contexts. First, although this study is focused on the general area of foreign language vocabulary, it does not cover all of the issues or processes involved in L2 vocabulary acquisition. This research is focused not only on deliberate vocabulary study and assessment, but more specifically on vocabulary learning through the use of traditional paper style word cards. Therefore, this study does not examine issues related to incidental vocabulary acquisition through methods such as extensive reading or listening. In addition, due to controlling various confounding variables, the treatment used in Experiment 1, and two of the treatments used in Experiment 2, are not as ecologically valid as they could have been; they do not accurately represent a real-world card making process.

Second, this study is limited in terms of participants and learning context. One delimitation is that the sampling techniques used limit the generalizability of the results and conclusions. The study uses a convenience sample consisting of participants from only one or two private Japanese university, whereas a bigger sample size from a variety of universities or educational institutions would have provided much more generalizable results. In addition, as most participants were drawn from a women's university, the sample consists of mainly female participants. Although the sample includes students from different departments who are at a variety of language proficiencies, and who have somewhat different learning backgrounds and goals, participants are also all aged between 18 and 22 years old. All of these factors severely limit the generalizability of the study. However, as discussed in the research design sections of this study, design choices have been made in order to limit the negative effects of using intact groups, for example, choosing within-subjects designs, and collecting qualitative data about participants.

In addition, while variables such as time-on-task, educational background, language proficiency, and study materials are controlled as far as possible, other variables such as language input from other sources, and number of self-study hours with other materials were difficult to obtain and could not be included in the statistical analysis. However, the qualitative data collection and analysis aimed to probe the issues, and these limitations are discussed and factored into the interpretations of the statistical data. Also, issues such as working memory, IQ, average GPA, and socio-economic status are not factored into the study.

Organization of the Study

In Chapter 2 I present a detailed review of the relevant foreign language vocabulary acquisition literature. I discuss various issues and theories relating to foreign language vocabulary acquisition, present findings in the field of L2 vocabulary research, and examine common pedagogical approaches to developing second language vocabulary proficiency. In the final part of Chapter 2, I present a further discussion of the current gaps in the L2 vocabulary literature that justify this research. In Chapter 3 I describe the methodology and results of the first experiment, which focuses on the learning outcomes of word card creation. This includes a description of research context, participants, treatments, materials, data collection, analysis methods, and results and discussion. In

Chapter 4 I describe the methodology and results of the second experiment, which compare the learning outcomes of three different approaches to word card study. In Chapter 5 I give a detailed description of the research design, methodology, results and discussions related utilized in the final set of experiments. These experiments examine various issues concerning testing of deliberately studied vocabulary. Then, the overall findings of the various experiments are discussed in Chapter 6. Finally, in Chapter 7, Conclusion, I summarize the main findings, discuss the limitations of the study, make suggestions for future research, and make concluding comments.

CHAPTER 2

REVIEW OF THE LITERATURE

Principles of Vocabulary Acquisition Pertinent to This Study

As mentioned above, there is no current or widely accepted model of second language vocabulary learning. As Tseng and Schmitt (2008) stated, "Given the complex nature of vocabulary knowledge, it is reasonable to assume that the process of learning this knowledge might have its own complexities. Thus, it is not surprising that there is currently no generally accepted theory of vocabulary acquisition…which has been able to capture all of the complexities of the acquisition process" (p. 358). However, there are a number of principles and underlying assumptions about vocabulary acquisition that have shaped the research questions and treatment designs in this study.

First, a key principle in SLA that is especially pertinent to vocabulary acquisition, is the notion of frequency or repetition, something Ellis (2002) has called *frequency effects*. Ellis and many others have stated that what is referred to as frequency, repetition, or number of meetings is strongly and positively correlated to vocabulary acquisition (Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014; Schmitt, 2008). Brown (2012) stated that "The frequency model of vocabulary learning, the idea that words are learnt broadly in order of their frequency, is routinely applied in language teaching, testing and research" (p. 21). The frequency model is directly connected with repetition as the more frequent a word is, the more times it will be met by a learner. While it is known that task complexity and levels of engagement or involvement are extremely important in vocabulary acquisition (Laufer & Rozovski-Roitblat, 2014), the number of meetings with a word is still an important determinant of vocabulary learning (Folse, 2011; Nation & Webb, 2011, pp. 42–43, 159; Schmitt, 2008).

The frequency model is strongly connected to having learners self-select words from frequency-based word lists. This is something that is an underlying factor in my classroom approach to deliberate vocabulary study and has necessitated the need for a suitable test system such as the peer-to-peer system investigated here. As word lists such as the General Service List (GSL), New General Service List (NGSL), the Academic Word List (AWL), and New Academic Word List (NAWL) are based on frequency counts, the words in these lists are likely the ones that learners will most frequently and repeatedly meet in reading and listening texts (Browne, 2014). The use of frequency-based word lists is also supported by the cost-benefit principle as described by Laufer (2014) who stated that "the cost-benefit principle in vocabulary learning states that learners should get the best return for their learning effort, i.e. learn words that they will meet and use often" (p. 39).

Recently, researchers such as Behzadi and Haji Pour Nezhad (2014), Laufer and Rozovski-Roitblat (2014), Nakata (2015), and Webb (2005, 2007, 2009) have shown that it is not only frequency that is important, but also what learners do with a word, with complex tasks yielding better results overall. Nevertheless, research has repeatedly supported Ellis's (2002) assertion that frequency and the amount of repetition have a strong effect on vocabulary acquisition (Barcroft, 2007a; Brown, 2012; Edy, 2014; Eckerth & Tavakoli, 2012).

The reason for the strong focus on deliberate vocabulary techniques in this study comes from the various research evidence showing that repeated exposure to vocabulary through incidental vocabulary study alone is not enough to make the sufficient gains needed (Elley, 1989; Hulstijn, 1992; Pellicer-Sánchez & Schmitt, 2010; Waring & Takaki, 2003). On the contrary, numerous researchers have shown the efficacy of deliberate vocabulary study (Cobb, 1999; Laufer & Rozovski-Roitblat, 2014; Nakata, 2015; Waring, 2004). These results were summarized by Schmitt (2008) when he stated, "It can no longer be assumed that an adequate lexis will simply be 'picked up' from exposure to language tasks focusing either on other linguistic aspects or on communication. Rather, a more proactive, principled approach needs to be taken in promoting vocabulary learning" (p. 333). As deliberate study offers the opportunity for multiple and frequent meetings with target vocabulary items in a time efficient manner, it too is supported by the frequency model (Brown, 2012; Ellis, 2002) and the cost-benefit principle (Laufer, 2014).

In this study, word cards were chosen as the deliberate vocabulary study method due to both theoretical benefits (Mondria & Mondria-de Vries, 1994; Nakata, 2020, p. 304), and empirical support (Komachali & Khodareza, 2012; Kuo & Ho, 2012; Reynolds & Shih, 2019; Waring, 1997). See Nation (2013) for a detailed review on the features, principles, and research relating to word cards (pp. 437–478). As discussed by Folse (2011, 2006) and Schmitt (2008), any activity that provides repeated meeting with vocabulary is beneficial, however, word card use has been repeatedly supported through the research noted above, and others (Emrah & Omur, 2010; Nakata, 2011; Nikoopour &

Kazemi, 2014; Schmitt, 1997, pp. 199–227). The success of word cards can be attributed to a number of factors described by Mondria and Mondria-de Vries (1994), Nakata (2008) and Nation (2013, p. 437–471). First, expanded spaced rehearsal can be implemented more easily as learners are able to divide cards into several decks, allowing them to review difficult or unknown items more frequently than easy or better-known items. Second, word cards allow for active recall of the L2 word form and its meaning separately because the L2 and L1 forms of the word are presented on different sides. Third, as words are presented separately and can be shuffled or studied in different orders, the chance of serial learning and list effect are removed, resulting in more opportunity for free recall practice, which in turn leads to better memory and recall ability later on. Furthermore, due to the relative efficiency and affordance for repeated, focused study of target items, their use is again supported by issues such as the frequency model and the cost-benefit principle. Therefore, word cards were used in this study due to their link with various theoretical underpinnings, and as a result of research-based results in support of them.

Overall, theories and principles relating to frequency, repetition, cost-benefit, and self-selection from frequency-based word lists have contributed to the overall design of the research questions, treatments, and materials used in this study. However, numerous issues affecting vocabulary acquisition have not been included in this study, for example, this study is not concerned with incidental learning, or vocabulary item factors such word length, orthographic and phonological regularity, and concreteness/abstractness.

The Relationship Between Vocabulary and Foreign Language Proficiency

As noted in Chapter 1, it is widely understood that vocabulary knowledge is directly connected to language use, and for many, vocabulary is viewed as the most important aspect of language proficiency because, as Wilkins (1972) stated, while without any grammatical knowledge very little can be conveyed or understood, without vocabulary knowledge nothing can be conveyed or understood. Whilst grammatical ability is very important and can greatly improve language accuracy, comprehension, and production, vocabulary knowledge is vital to understanding spoken or written texts, and in conveying ideas and meaning through writing and speaking (Laufer & Shmueli, 1997; Nation, 2013, pp. 56–56, 216–218; Schmitt, 2008; Qian, 1999).

Research has indicated that learners need to understand thousands of words in a written or spoken text if they are to gain reasonable comprehension (Hu & Nation, 2000; Laufer, 1989; Nation & Waring, 1997, pp. 6–19; Schmitt, Cobb, Horst, & Schmitt, 2017). While the exact amount of text coverage needed for successful reading or listening comprehension and spoken or written communication is still unclear, it is widely believed that in most contexts 95% to 98% is the minimum for good comprehension levels (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; van Zeeland & Schmitt, 2012). In order to reach this level, learners need knowledge of not only thousands of words, but thousands of word families (Hazenberg & Hulstijn, 1996; Hirsh & Nation 1992; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010). A word family includes “the base word, all of its inflections, and its common derivatives” (Schmitt, 2000, p. 2). For example, *respect*, *respects*, *respectful*, *respective* and *respectively* are all included in one word-family. In terms of the number of words learners need, Laufer and Ravenhorst-Kalovski (2010) state that to understand 95% of the words in a text, 4,000-5,000 word-families are needed, but to reach 98% coverage can require 8,000 word-families. In another study, Nation (2006) found that if a learner wants to attain the 98% coverage needed for dealing with written text, 8,000-9,000 word-families are needed, and 6,000-7,000 word-families are needed to deal with spoken texts. However, others have concluded that while 98% of lexical knowledge gives learners a higher level of comprehension, 95% coverage might be enough for listening comprehension, and only 3,000 word-families might be enough (van Zeeland & Schmitt, 2012; Schmitt, 2008; Schmitt et al., 2017). Therefore, while the exact number is not known, it seems that between 2,000 and 7,000 word families are needed if one is to take part in successful spoken discourse in English, and 8,000 to 9,000 word families are needed for successful written communication (Laufer & Ravenhorst-Kalovski, 2010; Schmitt, 2008; van Zeeland & Schmitt, 2012).

The importance of vocabulary is demonstrated not only by the above-mentioned studies, but by the growing body of research showing a positive correlation between vocabulary size and L2 reading ability (Anderson & Freebody, 1981, pp. 77–171; Laufer, 1992a, 1992b, pp. 126–132; Nation, 2006; Qian, 1999, 2002). Other researchers such as Astika (1993), Engber (1995), Laufer and Nation, (1995), and Olinghouse and Leaird (2009) have reported on the positive relationship between vocabulary knowledge and writing skills, with larger vocabulary sizes being positively correlated to quality of writing. It also follows that it is impossible to express oneself clearly in spoken form without an adequate vocabulary (Carter & McCarthy, 2014, pp. 202–204; Webb, 2009). Schmitt (2008) reported on the importance of vocabulary knowledge for speaking and writing skills in his review of numerous recent vocabulary research papers.

As seen above, while the exact number of words needed for successful foreign language use is not known, it is clear that learners need knowledge of many thousands of word families for success in reading, writing, speaking and listening. Without enough vocabulary knowledge, comprehension of written and spoken texts will be impossible, and written and spoken communication will break down. Therefore, it is clear that learners must devote considerable time and effort to learning second or foreign language vocabulary, and language educators have a responsibility to facilitate this as much as possible (Folse, 2011; Nation, 2013, pp. 2–3, 589–590; Schmitt, 2008). As Schmitt stated, "One thing that students, teachers, materials writers, and researchers can all agree upon is that learning vocabulary is an essential part of mastering a second language" (p. 329). While there is currently no clear model or widely accepted theory of foreign language vocabulary acquisition, and numerous studies claim the superiority of one study method over others, there are a number of widely accepted vocabulary methodologies available to foreign language vocabulary learners, all of which fall into one of two main approaches; incidental learning and intentional learning.

Incidental and Intentional Vocabulary Learning Approaches

There are two main approaches to learning foreign language vocabulary; incidental vocabulary learning and intentional vocabulary learning (Hulstijn, 2001, p.

271; Nation, 2013, p. 92–93; Schmitt, 2008), both of which have received considerable attention in terms of research and discussion in the academic literature. As discussed below, because each of these approaches have been shown to be both effective and problematic in their own right, the current view widely adopted by researchers and teachers is that both are necessary for learners to acquire the many thousands of words needed for foreign language comprehension (reading and listening) and production (writing and speaking) (Hirsh & Nation, 1992; Nation, 2013, pp. 59, 348; Schmitt, 2008; Webb, 2020, pp. 6–7).

Incidental vocabulary learning is also often referred to as implicit or indirect vocabulary learning and is understood to be the process of acquiring vocabulary while engaging in other activities, most commonly while reading, but also through speaking and writing activities. There have been a number of definitions of incidental vocabulary learning, for example, Hulstijn (2001) describes it as “the learning of vocabulary as the by-product of any activity not explicitly geared to vocabulary learning” (p. 271). Barcroft (2015) gives a similar definition when he states that “Incidental vocabulary learning refers to acquiring new words from context without intending to do so, such as when engaging in a conversation or reading a text for meaning and processing new words as input and inferring their meanings” (p. 25). For example, new vocabulary or more elaborate knowledge of previously studied words can be incidentally learned through reading a book, watching a movie, or listening to music.

This type of vocabulary learning approach has many supporters, some of whom think it is the singularly best method of learning vocabulary (Elley, 1989, 1991; Krashen, 1989). Others, such as Huckin and Coady (1999) offer support for incidental vocabulary learning when they state that “except for the first few thousand most common words, vocabulary learning predominantly occurs through extensive reading, with the learner guessing at the meaning of unknown words” (p. 182). However, the limitations of incidental vocabulary learning cannot be ignored. Incidental vocabulary learning has been shown to be not so effective for learning many new words, but it is very good for increasing knowledge of known words. During incidental vocabulary learning, learners see or listen to the words used in context, so they can acquire more knowledge about these items, for example, meaning, spelling, how they are used in a sentence, and pronunciation (Nation, 2013, pp. 355–360). Therefore, several researchers suggest that incidental vocabulary learning is especially effective for extending learner’s vocabulary knowledge (Day, Omura, & Hiramatsu, 1991; Folse, 2011; Vidal, 2003). In addition, it is argued that incidental learning is more interesting for learners than deliberate learning because it provides a context for the target items and usually involves self-chosen, easy activities that are supposed to be pleasurable, for example, extensive reading (Day, Omura, & Hiramatsu, 1991; Folse, 2011). Therefore, it is likely to be more motivating and easier to do for extended periods.

The other approach to vocabulary learning is intentional vocabulary learning, which is also often termed as deliberate or explicit vocabulary learning, and sometimes as language-focused learning (Nation, 2013, p. 217). Intentional vocabulary study involves learners carrying out activities with the sole purpose of acquiring vocabulary, and includes activities such as studying from lists, word cards, or from dictionaries (Folse, 2011; Luppescu & Day, 1993; Webb, 2020, p. 5). Hulstijn (2001) defined intentional vocabulary learning as “any activity aiming at committing lexical information to memory” (p. 271). Nation (2013) similarly states that, “Intentional learning occurs when a learner deliberately decides to learn a particular word or group of words and focuses on this learning” (p. 307). Barcroft (2015) provided another helpful definition when he stated that intentional vocabulary learning “refers to learning new words while consciously attempting to do so, such as when studying a list of new words, trying to learn new words while viewing word-picture pairs, or consciously attempting to learn new words from context while reading a text” (p. 25).

Many researchers have shown deliberate vocabulary learning to be a very effective method of acquiring new words relatively quickly (Elgort, 2011; Hirschel & Fritz, 2013; Komachali & Khodareza, 2012; Waring, 2004). In addition to the above-mentioned research, many other researchers and vocabulary experts support deliberate vocabulary study as a quick and efficient method of learning many new words (see Nation & Webb, 2011, pp. 1–59) for a comprehensive review of such studies). Therefore, given the great number of words learners need to know, and given the shortcomings of incidental vocabulary, learners should carry out significant amounts of regular and repeated deliberate vocabulary study.

Numerous researchers on deliberate or intentional vocabulary study support the above view, and many results have shown it can be an effective way to make substantial vocabulary gains in relatively short periods of time (Cobb, 1999; Laufer & Rozovski-Roitblat, 2014; Mondria, & Wiersma, 2004, pp. 79–100; Nakata, 2008). In addition, given the importance of repetition (Nation, 2013, pp. 107–110), repeated deliberate vocabulary study can be seen as essential for learners to make significant improvements in terms of vocabulary size (Folse, 2006; Schmitt, 2008; Webb, 2005). However, which method of deliberate study is most effective is still an issue that is widely discussed, and one that has by no means been settled by empirical evidence. For example, some researchers have pointed to vocabulary notebooks as an effective deliberate vocabulary study technique (Schmitt & Schmitt, 1995; Walters & Bozkurt, 2009), while others have highlighted significant outcomes of learning from word lists (Kitano & Chiba, 2019; Nakata, 2008). In addition, many L2 vocabulary researchers have argued in favour of word cards (Laufer, 2003; Mondria & Mondria-de Vries, 1994; Schmitt, 1997, pp. 199–227; Waring, 2004). Furthermore, the efficacy of computerized or digital vocabulary study techniques has also been noted by Dodigovic (2013), Ellis (1995) and Nakata (2008, 2020), all of whom stated that digital methods might offer more affective benefits than traditional, paper-based methods. This viewpoint has recently been supported by researchers such as Cennet (2011), Dodigovic (2013), Emrah and Omur (2010), and Nikoopour and Kazemi (2014). Therefore, while it is clear that deliberate vocabulary study is a necessity for foreign language learners, there is no clear consensus on which method is best, or even which features of each study method contribute the most to vocabulary acquisition.

However, while deliberate vocabulary learning is an effective method for quickly learning many new words, it has its own problems. For example, it can be boring and demotivating for learners (Reynolds & Shih, 2019; Tanaka, 2017; Wilkinson, 2015). This is a serious problem because motivation is widely viewed as a key determinant of success in language learning (Dörnyei, 2001, p. 1; Schmitt, 2008). Therefore, as discussed below, many researchers and experts believe it is important to also do more enjoyable vocabulary incidental vocabulary learning activities in addition to deliberate study.

The Need for Both Intentional and Incidental Vocabulary Learning

Just as there is no one best method or approach to facilitate successful language learning, there is no single method for acquiring vocabulary (Nation, 2103; Schmitt, 2008). Due to the limitations or problems associated with each method discussed above and given the numerous research results generated over the years, a combination of each type of vocabulary learning; intentional and incidental is deemed necessary (Folse, 2011, 2004; Laufer, 2003; Nation, 2013, pp. 3, 59, 348; Webb, 2020, pp. 4–5). Schmitt (2008) best summarized the current view of many leading vocabulary acquisition researchers with regards to the two forms of learning by stating that it is "clear that intentional and incidental approaches are not only complementary, but positively require each other" (p. 353). Webb (2020) makes a similar point when he states that the two approaches are not in competition with other but serve as useful compliments to one another (p. 6).

One of the reasons for the combined approach is that, as mentioned, although a great deal of research has been focused on the benefits of incidental vocabulary learning, especially from reading, many researchers report quite small vocabulary gains (Elley, 1989, 1991; Elley & Mangubhai, 1983; Laufer, 2003; Waring & Takaki, 2003). In addition, another problem connected to incidental vocabulary learning is that when using incidental vocabulary learning strategies, for example, guessing meaning from context, there is a possibility that learners will misunderstand word meanings, but this is not the case with intentional study techniques such as checking meanings in a dictionary (Bensoussan & Laufer, 1984; Mondria & Boer, 1991). However, many incidental vocabulary learning activities, such as extensive reading and extensive listening, offer other linguistic benefits such as fluency development, schema building, and vocabulary consolidation. As such, while these activities are commonly accepted as sound pedagogical activities offering a wide range of benefits, from a purely vocabulary acquisition point of view, they do not provide great enough learning outcomes to enable learners to acquire the thousands of words they need for comprehension and communication (Laufer, 2003; Laufer & Rozovski-Roitblat, 2014; Zhang & Lu, 2014).

Therefore, if students are to obtain an understanding of the large number of words needed for successful and meaningful communication, as it is most likely impossible to do so solely from incidental learning, a significant amount of deliberate study is needed (Folse, 2011; Elgort, 2011; Nation, 2013, pp. 59, 348; Schmitt, 2008). However, given the often times boring or demotivating nature of deliberate vocabulary study (Reynolds & Shih, 2019; Tanaka, 2017; Wilkinson, 2017), carrying out other more fun and enjoyable incidental activities that build on previously learned word knowledge is highly recommended.

A further point of interest here is that, as noted by many authors, currently most EFL vocabulary components of a curriculum, whether deliberate or incidental, are focused on building receptive vocabulary knowledge (Schmitt, 2008; Stewart, 2012;

Webb, 2009). Therefore, it is generally agreed that foreign language learners' receptive vocabulary knowledge is much greater than their productive knowledge (Mondria & Wiersma, 2004, pp. 79–100; Schmitt, 2008; Webb, 2005). However, while receptive knowledge is important for reading and listening ability, productive knowledge of thousands of words is needed for spoken and written communication (Nation, 2006; Schmitt, 2008; Webb, 2005, 2009). Therefore, a four-strands approach that utilizes a combination of both deliberate and incidental vocabulary learning activities involving reading, writing, listening and speaking activities offers the highest chance of success (Nation, 2007, Webb, 2020, p. 8).

In conclusion, as noted above, both intentional and incidental vocabulary learning offer their own distinct benefits and weaknesses but utilizing both methods can maximize a learner's vocabulary development. Deliberate vocabulary learning can help learners build their vocabulary very quickly but done on its own could lead to a lack of motivation for vocabulary study. Incidental learning does not offer enough vocabulary learning outcomes to be the sole method used, especially for low level learners, but can be an enjoyable way of learning new knowledge about previously studied words while improving other skills such as reading, writing, listening and speaking. Therefore, language curriculum designers and teachers should try to offer courses that give learners the chance to utilize both of these complimentary approaches.

The Involvement Load Hypothesis

It might seem like common sense that the more a learner is engaged or involved with a new word, the more likely it is that they will learn it. However, the terms engagement and involvement are vague and relatively unhelpful as pedagogical terms (Sarbazi, 2014). The use of these terms alone does not give teachers or researchers any concrete method or rubric to assess the probable efficacy of a vocabulary learning task. Therefore, there have been numerous attempts at defining this notion more precisely, with one of the earliest and most famous being Craik and Lockhart's (1972) *Depth/Levels of Processing Hypothesis*. This hypothesis is often referred to as the depth of processing model and states that the more attention given to an item, and the more manipulation involved with the item, the greater the chances it will be remembered. They suggested that long term memory retention depends on how deeply information is processed during learning; information that is processed deeply stays in memory longer than that which goes through a shallower processing (Tsubaki, 2012). As an example, if learners are given a new L2 target word alongside their L1 translation, this would involve shallower processing than encountering the new L2 word, then looking the meaning up in a dictionary. Following on from that, under the hypothesis, the student could then go through a much deeper process by, for example, creating a word card for that item. If a learner was to then add additional information to their word card such as pronunciation information, other word forms, and example sentences, each additional type of information combined with the time-on-task and actions involved would represent a deeper processing level, and in turn would produce a more robust memory trace.

As a result of criticisms of the depth of processing model, mainly that the definition of depth of processing was unclear, the model was oversimplified, and that the memory of words is influenced by other factors not included in the model (Baddeley, 1999), a new model was sought. Therefore, Craik and Tulving (1975) expanded on the theory by adding the notion of elaboration. They proposed that when new information is connected to preexisting information, it is enriched and more likely to be remembered. However, the main criticism of Craik and Tulving's proposal was, like that of Craik & Lockhart (1972), it still had an unclear definition of depth of processing. In addition, despite these two theories, it was difficult to operationalize the concept of involvement (Laufer & Hulstijn, 2001), meaning that it was difficult to use the theories to assess various tasks in terms of their probable success for vocabulary learning.

As a result, using the depth of processing theories described above as the starting point, *The Involvement Load Hypothesis* (Laufer & Hulstijn, 2001) was developed. The authors claim that retention of unfamiliar words is contingent upon the involvement load of a task, a construct the authors call *task-induced involvement*. The greater the amount of involvement needed to complete a task, the greater the retention of unfamiliar vocabulary will be. Under the Involvement Load Hypothesis (ILH), task-induced involvement includes both cognitive and motivational factors, specifically the amount of need (motivational factor), search, and evaluation (cognitive factors) that a task induces. The inclusion of motivational factors as well as cognitive factors was one of the attractive points of the ILH as it is well documented that motivation plays an important role in language learning and can be just as influential as cognitive factors (Tsubaki, 2012). In addition, the ILH offers researchers and teachers a more operationalized description of involvement and provides a clear scoring rubric that can be used to assess the potential strengths of various vocabulary activities.

Three Components of Vocabulary Learning in the Involvement Load Hypothesis

As mentioned, Laufer and Hulstijn (2001) provided three components of task-induced involvement: need, search, and evaluation. Each of these components was assigned up to three degrees of value; none (0), moderate (1), and strong (2). Also, as seen in Table 2 below, an explanation was provided for each degree of the three components in order to allow researchers and teachers to assess the task-induced involvement load of various vocabulary learning tasks.

The first component, *need*, refers to the motivation to learn a target word, be it an internal or external need. If the learner does not need to understand or use the target word in a task, then the score is 0. However, if there is an externally imposed need to learn or use the word, then the score is 1. Finally, if there is an internally felt need, then the score is 2. An example of where there is no need, would be when a learner encounters an unknown word in a reading passage but decides it is not important to the comprehension or enjoyment of the passage (Nation & Webb, 2011, 2–7). An example of an external need with a score of 1 would be when a student is instructed to learn the meaning of new words by the teacher, possibly in order to later pass a test in class. If there is an internal need to learn or use the word, or the learner autonomously decides they need to learn the word, then the score is 2. An example of an internal need would be when a learner encounters an unknown word during pleasure reading and wants to learn its meaning in order to better understand the story they are reading.

Table 2. *The Degrees of the Involvement Load Hypothesis Components*

Component	Degree of Involvement Load	Explanation
Need	0 (None)	Learner does not feel need to learn the word
	1 (Moderate)	Learner is required to learn the word
	2 (Strong)	The learner decides to learn the word
Search	0 (None)	Learner does not need to learn the meaning or form of the word
	1 (Moderate)	Learner finds the meaning of the word
	2 (Strong)	Learner finds the form and meaning of the word
Evaluation	0 (None)	Learner does not compare the word with other words
	1 (Moderate)	Learner compares the word with other words in provided context
	2 (Strong)	Learner compares the word with other words in self-provided context

Note. Adapted from Tsubaki (2012).

The second component is s*earch*, and this is said to occur when a learner has to find the meanings of target words. While Laufer & Hulstijn (2001) originally decided that scoring for search could be only 0 or 1, according to Nation and Webb (2011, p. 4), Laufer has now acknowledged that a three-level scoring rubric is more suitable. An example of a condition with no search would be if a learner encounters an unknown L2 word in a reading passage but decides it is not important to understand the passage and therefore makes no attempt to guess their L1 meaning or look the word up in a dictionary. However, if the learner stops reading to consult a dictionary to find their L1 meaning, then search has taken place. This process is known as receptive retrieval (Nation & Webb, 2011, p. 4) and receives a score of 1 for search. However, if the learner engages in productive retrieval, which occurs when a productive activity requires the learner to retrieve or look for the L2 form of a word, the score would be 2. This type of activity is less commonly discussed in the literature, but an example could be when a learner does not know the L2 word needed to express an L1 concept, for example, in a writing task, and therefore enters the L1 word into a bilingual dictionary and retrieves the L2 form of the word before using it in their sentence.

Finally, *evaluation* involves comparison of a target word, or information about a word, with other words. Nation and Webb (2011, p. 4) provided an alternative, and possibly more helpful, description of the evaluation component by stating it involves making choices between options or providing a suitable context in which the word can be used. The decision about whether a word fits or not or is the most suitable choice could be based on various issues including grammar, register, context, or nuisances of meaning. If a learner does not need to decide which word or sense of the word to use, then the score would be 0, but if they do, the score would be 1. To achieve a score of 2, a learner would have to use the word in context and provide that context themselves. Common activities involving a low level of evaluation require learners to compare the target word with its context of use, and decide if it fits or not, or is the best choice (Sarbazi, 2014; Tsubaki, 2012). Example activities include fill in the blank activities, matching exercises, and choosing the best word from multiple choices, all of which would induce an evaluation score of 1. Activities with a strong evaluation index include writing original sentences using the words, writing essays or other compositions, or using the target items in a speaking task.

Common Application of the ILH

The ILH, its three components, and the degrees of involvement for each, have provided researchers and teachers with a much clearer and easier method of assessing or predicting the efficacy of various vocabulary learning activities. However, whether due to the authors' original intention or not, the ILH, has almost solely been used as a method of assessing incidental vocabulary learning tasks such as reading, listening and writing. However, there is no clear reason why the theory itself, or its scoring index, cannot be used to assess or predict the efficacy of various deliberate study techniques, including word cards. In fact, Nation and Webb (2011) did exactly that in the table at the end of their book (p. 318–319).

Research in Support of the ILH

Laufer and Hulstijn (2001) note that a great deal of support for the ILH predates its formulation through various studies examining the effect of different task types. While not the primary focus, many studies comparing the effect of different task types have included tasks with varying degrees of task-induced involvement, for example, Hulstijn (1992), Hulstijn, Hollander, and Greidanus (1996), Nagy, Herman, and Anderson (1985), and Watanabe (1997). Others such as Cobb (1999), Mizumoto (2013), Newton (2013), and Webb (2005, 2007) have examined the effects of varying degrees of engagement, variables that are very similar to the involvement load constructs. Generally, activities with higher levels of task-induced involvement have been shown to produce better learning outcomes. However, as many of the studies above did not specifically focus on the ILH; they did not compare treatments with different scores on the ILH constructs, they cannot truly be used in support of the ILH.

However, some studies directly assessing the effects of tasks with differing task-induced involvement loads have been carried out. The first of such studies was carried out by Laufer and Hulstijn (2001) who conducted two parallel experiments in two countries. Advanced Dutch- and Hebrew English learners were formed into six intact groups and retention of ten unfamiliar words encountered in incidental learning tasks was investigated across three tasks types. Task 1 included reading comprehension with marginal glosses, Task 2 included reading comprehension plus gap fill activities, and Task 3 was a composition writing task requiring the use of target words. The tasks were designed to ensure they each had different task-induced involvement loads; differing levels of need, search and evaluation. The result indicated that Task 3, which had the highest involvement load score, led to better retention than Task 1 and Task 2, thus providing support for ILH.

In a more recent study, Keating (2008) used three tasks to predict the vocabulary retention effects of tasks with different task-induced involvement loads. Seventy-nine beginning learners of Spanish completed one of three vocabulary learning tasks that varied in the amount of involvement, something Keating also referred to as mental effort. As nonsense words were used, there was no need for a pretest, but passive and active knowledge of the target words was assessed immediately after treatment, and a delayed posttest was given two weeks later. The first task involved a reading passage with marginal glosses (weak involvement), task 2 involved a reading passage with a gap fill activity (moderate involvement), and task 3 required participants to write original sentences using the eight target words (strong involvement). In Task 1, which had an involvement load index of 1, the participants simply had to read a passage and use the target words and their glossed meanings to answer five true/false comprehension questions. The involvement load index was 1. Group 2 read the same text but with the words deleted from the passage, then using a brief definition, an example sentence and an L1 gloss for each target word, participants were instructed to fill in the blanks with the correct word. The involvement index was 2. Participants in the third group were provided with the same information for each target word, but did not see them in a text, and were required to write original sentences with the words. The involvement load index for this task was assessed to be 3. Based on the ILH index scores, it was predicted that group 3 would perform the best, followed by group 2, and that group 1 would perform the worst. Keating stated that results were in line with the predictions of the Involvement Load Hypothesis because retention was highest in the sentence-writing task, lower in the reading plus fill-in task, and lowest in the reading comprehension task. However, the results did not hold true when time-on-task was factored in, with the benefit associated with more involving tasks fading away. Nevertheless, Keating concluded that the results strongly supported the ILH, and it can be used to predict the efficacy of vocabulary learning tasks and to support their inclusion in foreign language curricula.

In a more recent study that included task-induced involvement load as an independent variable, Eckerth and Tavakoli (2012) investigated the differential effects of elaboration of word processing (involvement), and word exposure frequency on initial word learning and retention. There were 30 participants who were enrolled in three parallel advanced pre-university EAP courses at the English Language Centre of a British university. Participants were aged between 22 and 27, and were mostly Asian, but some had Slavic and Romance L1 backgrounds. The data were analyzed using a 3 × 2 × 2 × 4 repeated measures MANOVA with task-induced involvement load (3 levels), exposure frequency (2 levels), and time of measurement (2 levels) as the independent variables, and word knowledge type (4 levels) as the dependent variable. The first independent variable involved three reading tasks that varied the task-induced involvement load (low in Task 1, medium in Task 2, high in Task 3) by imposing different reading conditions on the participants (reading a text with marginal glosses in Task 1, filling in text gaps from a word list in Task 2, reading a text with marginal glosses and writing a related composition in Task 3) (p. 233). The second independent variable, exposure frequency, was operationalized at two levels; one occurrence, and five occurrences. This involved presenting 10 target words in each reading text; five occurred once and five occurred five times. A pretest-posttest design was used with the dependent variable (word knowledge) tested through three weekly immediate posttests after completion of the tasks, and one delayed posttest three weeks after the last task was completed. The tests measured the participants’ receptive and productive word recall and word recognition and represented four dependent measures. All participants were exposed to all target words, all three reading texts, and the three reading conditions. The results of the repeated measures MAOVA analysis showed significant main effects for task-induced involvement load, $F(2, 27) = 30.98$, $p = .001$, $\eta^2 = .56$, exposure frequency, $F(2, 27) = 12.73$, $p = .001$, $\eta^2 = .305$, time of measurement, $F(2, 27) = 106.83$, $p = .001$, $\eta^2 = .79$, and word knowledge type, $F(2, 27) = 138.44$, $p = .001$, $\eta^2 = .83$. The results showed that a considerable amount of the variance in the participants’ vocabulary gains resulted from the effects of task-induced involvement load and word frequency, but the effects differed significantly over time and across word knowledge types. To look at the specific results in more detail, a series of repeated measures MANOVAs was carried out. The results showed strong effects for task-induced involvement load, $F(2, 27) = 41.71$, $p = .001$, $\eta^2 = .59$, and for exposure frequency, $F(2, 27) = 166.90$, $p = .001$, $\eta^2 = .85$, on initial word learning. The interaction between the two variables was also significant, with increased frequency and a higher involvement load improving word learning. However, long-term word retention was more dependent on task-induced involvement load than on word frequency. Tasks that involved elaborate processing of the form-meaning relationships resulted in more positive long-term vocabulary gains than simpler tasks that provided more meetings with the words. The researchers concluded that after first meeting target words, repeatedly carrying out vocabulary focused activities that increase involvement is superior to carrying out further tasks that do no more than provide repetition of target items, for example, repeated reading or listening. Eckerth and Tavakoli acknowledged that the number of meetings with a word has a significant effect on vocabulary learning, but they state that their results showed that receptive tasks are less efficient than more complex involvement-inducing tasks. Examples of such tasks include searching for words needed to complete activities, and productive activities such as gap fills, sentence writing, and speaking activities.

This was a well-designed study, however, one limitation arose from the small sample size. Nevertheless, the results lend support for not only the importance of repeatedly meeting target words, but also for including tasks in language courses that enable learners to carry out higher levels of involvement and processing such as using dictionaries, answering vocabulary focused questions, or writing sentences. While this study was focused on incidental vocabulary learning, the results regarding issues of repeated exposure to vocabulary and level of processing can clearly be applied to deliberate vocabulary learning techniques and practices, especially as some of the tasks used in the study could be seen as having an intentional vocabulary study focus.

In a very recent study, Laufer and Rozovski-Roitblat (2014) examined the effect of tasks with different levels of involvement, number of encounters with target words, and the combination of each on incidental L2 word learning over time. The participants were 185 EFL learners in 10 intact parallel classes. The participants had all studied English for 7 or 8 years, but they had not achieved the level needed for regular university courses, therefore, they were enrolled in pre-academic English courses. The independent variables were task type and number of encounters, and the dependent measures were scores on various delayed posttests that measured passive and active recall, and passive and active recognition. Data analyses were carried out using a repeated measures MANOVA in order to assess the vocabulary learning effects of (a) three task types (reading only, reading with a dictionary, reading and word focused exercises), (b) three number-of-encounters conditions, and (c) the interaction between task type and number of encounters on four types of word knowledge. In addition, an ANOVA was used to measure the differences between task type in each number-of-encounters condition separately, and a repeated measures ANOVA was used to measure the differences between the number of encounters conditions in each task. Additionally, the researchers used post-hoc Tukey Tests to analyze data pairs when the ANOVA results showed significant differences. However, the statistical data needed for expected statistical reporting are only available in an online appendix, but this is no longer available, so traditional statistical reporting of results is currently not possible.

The main results were that reading with word-focused exercises yielded the best results on all types of word knowledge. Reading with one exposure and subsequent focus on forms yielded the highest retention scores and required a considerably smaller number of encounters than reading only. The number of encounters had different effects in different task conditions, but overall, the greater the number of encounters, the more positive the effects. The researchers stated that there was a clear advantage of tasks that included vocabulary exercises over reading activities, irrespective of the number of encounters with the new words and the type of word knowledge. The researchers stated that the number of encounters was less important than the task, as the reading with a dictionary, and reading-focused and word-focused exercises resulted in a greater effect of repetition than the reading only condition. The main result of interest was that vocabulary-focused activities with a minimal number of encounters resulted in greater vocabulary gains than the reading activities with the maximum number of encounters. The researchers concluded that what learners do with the word might be more important than how many times they encounter it. They feel that activities that require higher levels of processing and a higher involvement load result in superior learning and retention outcomes. Overall, the results of this carefully carried out study offer further support for the involvement load hypothesis and strengthen the justification for the inclusion of deliberate vocabulary study activities in language curriculums, especially those that offer higher levels of processing and involvement.

In addition to the above, other studies showing superior vocabulary gains from vocabulary learning tasks with higher levels of the ILH constructs of need, search and evaluation include Pichette et al., (2011), and Rott, & Williams (2003). The main results of the above studies in terms of engagement or involvement load and self-regulation are summarized in Table 3.

Table 3. *Vocabulary and Involvement Load*

Engagement and Involvement Load	
Increased engagement/involvement leads to superior learning outcomes	Cobb (1999); Eckerth & Tavakoli (2012); Laufer & Rozovski-Roitblat (2014); Mizumoto (2013); Newton (2013); Pichette et al., (2011),
What learners do with words might be more important than how many times they meet it.	Eckerth & Tavakoli (2012); Laufer & Rozovski-Roitblat (2014); Pichette et al., (2011); Rott & Williams (2003)
Self-Regulation	
SR results in superior learning outcomes	Cobb (1999); Mizumoto (2013)

In summary, the combined results of the studies reviewed above show that the level of engagement or involvement that learners have with target vocabulary items is an important factor in second language vocabulary acquisition. Tasks that require higher levels of task-induced involvement are those that trigger multiple cognitive processes and involve working with words in different or more complex ways, and these types of activities have proven to be more beneficial to vocabulary learning than more simple tasks such as simply encountering words in reading passages. In addition, more self-regulated activities that require learners to reflect on their choices, or make individual decisions regarding which words to study, and how to study them, can also increase vocabulary acquisition.

Deliberate Vocabulary Study From Word Cards

Many vocabulary researchers support the use of word cards (Laufer, 2003; Mondria & Mondria-de Vries, 1994; Nation, 2013, p. 437; Schmitt, 1997, pp. 199–227), however, Paul Nation is probably the most well-known and ardent supporter of word cards for repeated deliberate vocabulary learning. Under the four-strands approach, this method falls under the language-focused learning strand, something often also referred to as deliberate vocabulary learning (Nation, 2013, p. 217). Vocabulary cards in their most basic form involve having the target language word written on one side, with the native language version on the other (Nation & Webb, 2011, p. 29). However, depending on learner or teacher preferences, other information such as part of speech, collocations, pronunciation information, or example sentences can also be included on the cards (Reynolds & Shih, 2019; Nation, 2013, p. 437).

In addition to the general support of word cards mentioned above, some research has been carried out in order to compare the effect of studying vocabulary from lists with that of studying from word cards. Generally, although learning from lists has proven to be a successful deliberate learning technique (Kitano & Chiba, 2019; Kuo & Ho, 2012; Nakata, 2008; Walters & Bozkurt, 2009), studying from word cards has been suggested to be a more effective method in terms of learning outcomes (Mondria & Mondria-de Vries, 1994; Nation, 2013, p. 451; Schmitt & Schmitt, 1995; Waring, 2004).

Research has suggested that the success of word cards can be attributed to a number of factors described by Mondria and Mondria-de Vries (1994), Nakata (2008) and Nation (2013, p. 437–471). First, by having the L1 and L2 forms of a word written on different sides, learners are given opportunities for retrieval; retrieving the meaning of the word from memory. Word lists or notebooks provide both the L1 and L2 information side-by-side, meaning that learners must cover up one of the forms in order to retrieve the other language meaning from memory, something that many students might not do. However, because the L2 and L1 forms are presented on different sides, when studying from word cards, learners are forced to try to remember and retrieve the word before then checking if they are correct, and this process is deemed to produce better learning outcomes (Nation, 2013, p. 451). Learners are essentially giving themselves a mini test each time they study from their word cards, and as a result, positive test effects will take place. In addition, word cards also afford a time efficient method for repeated rehearsal because students can quickly study the words again and again without the use of any other study aids. In addition, expanded spaced rehearsal, something that Nation (2013, p. 451) referred to as the spacing of repetitions, can be implemented more easily than when using notebooks or word pair lists, as learners are able to divide cards into several decks, allowing them to review difficult or unknown items more frequently than easy or better known items. Finally, if cards are shuffled or studied in different orders, there is little chance of serial learning and more opportunity for free recall practice, which in turn leads to better memory.

Mondria and Mondria-de Vries (1994) added further explanation for the efficacy of word cards in a comprehensive review of the theoretical and practical literature in support of the use of word cards. The authors discuss other positive aspects underlying the rationale behind the use of word cards for the series of studies carried out in this study. First, the authors pointed out that the word card system is flexible in terms of content; a variety of information such as translations, synonyms, language functions, and pronunciation can be added to the cards. Second, the authors agree with Nakata (2008) and Nation (2013, p. 451) by also stating that word cards are superior to learning from lists or textbooks because learners can separate their packs in order to focus on more unknown or new words, spending less time on easier or more known words that would be presented together on a list. This strategy, referred to as expanded spaced retrieval, is supported by various models such as Pimsleur's (1967) graduated interval recall model and Landauer and Bjork's (1978, pp. 625–632) expanding rehearsal strategy. The authors also pointed out that further support for word cards comes from evidence indicating positive learning effects of writing down the foreign language word on cards, something also shown more recently in Webb (2005). Taken together, this paper, along with the related literature discussed above, provides theoretical support for the efficacy of word cards as a deliberate study method or technique. However, theoretical support alone is not enough, therefore, a number of empirical studies concerned with the efficacy of word cards as a deliberate study technique are summarized and critiqued below.

In one of the first studies to experimentally investigate the efficacy of word cards, Waring (1997) investigated the required study time, and learning and retention outcomes of learning from word cards receptively and productively. The terms receptive and productive have different meaning to those in Webb's (2005) study. In this study, receptive learning referred to first seeing words in the L2 before checking the L1 equivalent, and productive learning means going from the L1 to the L2. The participants were 70 female Japanese learners of English aged between 18 and 20, taken from two intact classes at a private university in Japan, however, due to a lack of posttest data, the final *N*-size was 60. All participants were familiar with word card learning as a deliberate vocabulary learning strategy. The participants were supplied with two sets of 15 words cards, each containing the L1 and L2 form of a low-frequency word on different sides. The participants then learned the word sets either receptively (L2-L1) or productively (L1-L2). In order to ensure that the learners studied the cards correctly, Waring stated that all the cards were carefully arranged in terms of which side was facing up, and that each set had instructions such as *Learn from English to Japanese.* The participants were given as much time as they needed, and the time taken to learn receptively and productively was recorded. The task was deemed as completed when the participants were successfully able to recall each set of words twice (receptively or productively) without mistakes, then after a ten-minute distractor activity, a productive and receptive translation test was administered. Then, three delayed posttests were administered one day, one week, and three months after the treatment, respectively.

The data were first analyzed using a 2 (word sets) x 2 (receptive first or productive first) x 2 (time taken to learn receptively and productively) repeated measures MANOVA with word sets and order as the within subject variables, and time taken as a between groups variable. No effect was found for word set, $F(1, 56) = 0.02$, $p = .96$, or order; $F(1, 56) = 0.28$, $p = .60$, but a main effect for time was found, with the productive learning task taking significantly (26.1%) longer, $F(1, 56) = 113.2$, $p = .0001$. In terms of learning outcomes, a 2 (set A vs set B) x 2 (order 1 vs order 2) x 2 (type of learning) MANOVA was carried out, and a significant main effect for type of learning was found, with receptive learning yielding higher test scores, $F(1, 56) = 115.2$, $p = .0001$. Further analyses were carried out in order to examine whether words learned productively are better remembered when tested productively or receptively, and whether receptively learned words better remembered when tested productively or receptively, however, no statistical significance analysis was carried out. The results showed that while the mean scores on the receptive immediate and delayed 1 to 3 posttests were only slightly higher for the participants who completed the receptive task, the results of the productive test showed that the participants who completed the productive task had considerably higher mean scores than those who completed the receptive task. However, the results also indicated that the productive knowledge also decayed at a faster rate.

Waring made a number of conclusions based on the results, the first being that word cards are a time efficient method of learning vocabulary. It took an average of 21 minutes 28 seconds to learn 30 new words, which equates to 84 words per hour for decontextualized learning assuming no effect for fatigue. Based on the three-month delayed posttest data, Waring went on to state that with no further study, 60% of the target words will still be able to be recalled receptively or productively. Another conclusion was that receptive learning might be superior to productive learning as it yielded significantly better results, it is easier than productive learning in that it takes less time to gain knowledge of target items, and receptive knowledge seems to suffer less decay than productive knowledge. A final conclusion was that knowledge from receptively learned words can be available productively and vice versa.

Overall the results of this study indicated that (a) word cards are an effective method of vocabulary study, (b) both receptive and productive learning result in significant vocabulary gains, (c) receptive learning takes less time and suffers from less decay, and (d) both types of study result in gains in both types of recall. This is one of the only studies, other than Nakata (2008), that has provided empirical support for the efficacy of word cards as a deliberate study technique. This study is important because it was well controlled, carried out in natural learning conditions, and vocabulary gains were tested longitudinally. In terms of applying the results to real-life learning contexts, as under normal word card study conditions, it is likely that learners learn both receptively and productively at different times, this study offers empirical support for the efficacy of word cards as a deliberate study technique because both methods were successful.

Another of the few studies specifically examining word cards was conducted by Komachali and Khodareza (2012), who investigated the effect of using vocabulary cards on vocabulary knowledge. The participants were 50 female Iranian pre-university learners taken from a larger population and who were randomly assigned into two homogeneous groups of 25 learners; they were matched on everything but the treatments. A pretest-posttest design was used, and 60 vocabulary items selected from the students' textbooks were used in the tests. A 60-item multiple-choice test was split into two equal halves based on odd and even numbers and used as the pretest and posttest. During the experiment, the comparison group received "routine instruction" (p. 138), but the authors did not state what form this took, while the experimental group received the vocabulary card treatment. The vocabulary cards used in the experimental treatment consisted of the L2 word, pronunciation information, and different parts of speech on one side, and synonyms and antonyms, dictionary example, learner's example, and L1 definitions on the other side. The participants studied the words in the different conditions over the course of the semester. The independent variable was treatment with two levels; the routine instruction group and the word card group, and the dependent variable was the scores on the multiple-choice pre and posttests.

Independent sample *t*-test results revealed that while there were no significant differences between the two groups on the pretest, $t(48) = 1.129$, $p = .26$, the vocabulary card group performed significantly better on the posttests, $t(48) = 9.679$, $p < .001$. In addition, paired sample *t*-test results showed that while the comparison group did not make any significant gains between pretest and posttest, $t(24) = 2.009$, $p = .056$, the word card group did, $t(24) = 18.751$, $p < .001$. The researchers concluded that word cards are an effective way to increase learners' vocabulary knowledge and might be better than more traditional teacher-fronted approaches. The researchers stated that using vocabulary cards as a class activity facilitates student involvement due to issues such as peer interaction and involvement, sharing answers, paying attention, and giving examples, and these actions directly increase vocabulary learning outcomes.

This study offers further support for the use of word cards, as the researchers showed that a significant amount of learning took place through the use of word cards, and the word card group outperformed the more traditional teacher-fronted method. As this study is one of very few that involved real word cards, and it was carried out longitudinally, the results are important to both researchers interested in deliberate vocabulary study, and more specifically in word cards. However, as little information was given about the comparison group, it is difficult to make strong assumptions based on comparisons of the two methods. For example, no information was given regarding what constituted *routine instruction*, so it is unknown what the comparison group did. In addition, as no information was given about how much time the students devoted to vocabulary study over the semester, it is difficult to say whether a significant amount of learning occurred (pedagogical effect size). Another limitation of this study concerns the test; no information is given about the answer choices, and although the word cards contained seven types of information, the students appear to have only been tested on form or meaning recognition. If more sensitive measures had been used, the results might have been very different.

Kuo and Ho (2012) conducted another of the few studies to comparatively investigate L2 vocabulary acquisition using word cards and other methods. The researchers primarily compared the learning effects of word cards with that of studying from a word list, and the effect of spaced versus massed practice on English vocabulary retention. They also explored learners' attitudes toward each strategy. The participants were 120 Taiwanese EFL ninth graders in four intact groups; card-spaced, card-massed, list-spaced, and list-massed. The participants received identical eight-week instruction on 120 target words presented on either word cards or word lists. The word cards had one target word, its part of speech, and a sample sentence on one side, and the other side was left blank for participants to write the L1 equivalent in class. The 120 word cards were divided into three packs and distributed to participants in the first, fourth, and seventh week of the course. Word lists were presented on five pieces of A4 paper, with their parts of speech, and sample sentences on each list, but the participants wrote L1 equivalents in class. No information was provided about when the word lists were handed out. During the treatment stage, an instructor taught the 120 target words in class over eight weeks (15 words per week), and this was followed by a weekly achievement quiz the next day. In addition, 15 minutes of in-class time per week was allocated for the word-card group to practice and share their experience of using the word card strategy, and for the word-list group to practice the word list strategy and memorize the target words on their own.

The independent variable was the treatment with four levels: card-spaced, card-massed, list-spaced, and list-massed, and the dependent measures were scores on pretests, immediate posttests, and delayed posttests that were administered just before, one week after, and five weeks after the experiment, respectively. The vocabulary test measured the participants' receptive vocabulary knowledge (translating English words into Chinese) and productive vocabulary knowledge (translating Chinese words into English) of the 120 target words, with 60 items measuring each type of knowledge. Two questionnaires and focused interviews were used to examine the participants' attitudes toward the word card study method versus that of using lists.

The researchers used a series of paired-sample *t*-tests, independent-samples *t*-tests, and an ANOVA to investigate the effects of word card versus word list strategy, and spaced practice versus massed practice on learner's vocabulary retention. First, the results of three independent sample *t*-tests comparing the results of the word card and word list groups showed that while there was no significant difference between the groups on the pretest, $t(118) = -.81$, $p = .42$, there was significant differences on the posttest, $t(118) = 2.72$, $p = .01$, and delayed posttest, $t(118) = 4.31$, $p < .01$, with the word card strategy yielding better performance on vocabulary retention. The word card strategy resulted in a mean difference of 11 more words being retrieved than with the word list strategy on the posttest, and the gap was even larger on the delayed posttest with the word card strategy leading to 17 or more words being retrieved. Next, in order to examine whether spaced practice results in significantly better EFL vocabulary retention than massed practice, further independent-samples *t*-tests were carried out. There were no significant mean differences between the groups on any of the tests; pretest: $t(118) = -.96$, $p = .34$, posttest: $t(118) = 1.01$, $p = .31$, delayed posttest: $t(118) = 1.01$, $p = .31$. Finally, the results of three one-way ANOVAs comparing the pretest, immediate posttest and delayed posttest scores among the four groups: card-spaced, card-massed, list-spaced, list-massed group were analyzed in order to examine if spaced practice significantly reduces the effectiveness of the word card and word list strategy. The results showed that there were no significant differences found on the pretest, $F(3, 116) = 0.68$, $p = .56$, or on the delayed posttest, $F(3, 116) = 6.54$, $p = .21$. The only significant difference was found on the immediate posttest, $F(3, 116) = 2.84$, $p = .04$, and when analyzed using LSD post hoc analysis, the researchers stated that the findings imply that spaced practice widened the positive gap between word card and word list strategies more than massed practice, and that the card-spaced group significantly outperformed both the list-spaced and list-massed groups for the posttest. The researchers state that the results show that word card learning is superior to list learning, and that spaced rehearsal is superior to massed practice, especially when studying from word cards.

Questionnaires and individual interview data indicated that 70% of the word-card group preferred the word card strategy to the word list strategy, and that word cards were viewed as being more convenient. In addition, over 95% of the word-card group reported that they often or sometimes engaged in retrieval activities, while only 66.7% of the word-list group covered target words or their L1 equivalent meanings with their hands to retrieve them.

The researchers concluded that the statistical results showed that word card study is more effective than word list study, especially when combined with spaced practice. They also stated that the qualitative data suggested that this could be due to issues such as ease and frequency of retrieval activities afforded by the card design, plus the fact that learners' positive perceptions of the card method can lead to greater engagement and intrinsic motivation. Based on the overall results, the researchers concluded, "the word card strategy with spaced practice is strongly recommended for EFL junior high school students who need effective strategies to facilitate their vocabulary retention" (p. 38).

In summary, the results of this study suggest that (a) word card learning is more effective than learning from lists, (b) spaced practice does not result in significantly better vocabulary retention than massed practice within one particular study method, but (c) spaced practice increases the positive gap between the word card and word list strategies, and (d) learners prefer word card study over list learning. Taken together, the results provide further support for word card use. In addition, as the study was well controlled and carried out longitudinally in a natural learning environment, the results are significant to researchers and teachers interested in deliberate vocabulary study methodology, especially those interested in learning with lists and word cards. It should be mentioned here that a similar study by Kitano and Chiba (2019) found lists to be more effective than word cards, opposite results to those found by Kuo and Ho reported above. However, Kitano and Chiba's study involved only 25 participants and no statistical analysis was carried out; only descriptive statistics were reported.

In another study concerning word cards, Nakata (2008) compared the vocabulary learning outcomes of studying with word lists, word cards, and computers in order to identify which method results in superior vocabulary learning. The participants were 226 Japanese first- and second-year male senior high school students aged 15-17. Although it is stated that there were no differences between the groups in terms of standardized English test scores, no details are given regarding their English proficiency level. The participants were assigned to one of three treatment conditions, List group, Card group, and PC group, and each group studied ten English words using the assigned treatment materials; lists, word cards, or computers. The participants completed a pretest, posttest, and delayed posttest administered four days after the treatment ended. The participants were required to translate the target L2 word into the L1, and all of the tests were the same except that the order of words was changed. The data were first analyzed using a two-way ANOVA with two independent variables of group (treatment) and time-on-task, and the dependent variable was the vocabulary tests scores. The results showed that the PC group had the highest scores among the three groups, but their average study time was also the longest, $F(2, 195) = 12.31$, $p < .001$. Therefore, because the PC group's superior posttest scores might have been a result of their longer study time, a one-way ANCOVA was used to control for the differences in the three groups' study time. For the ANCOVA, the independent variable was group (treatment), the dependent variable was the vocabulary tests scores, and the covariate was time-on-task. The ANCOVA results showed that when time-on-task was controlled, while there was no significant difference in the immediate posttest scores of the three groups, $F(2, 194) = 0.39$, $p > .05$, differences in the delayed posttest scores were statistically significant, $F(2, 180) = 3.23$, $p < .05$. Bonferroni post hoc analysis showed that the PC group outperformed the list group ($p < .025$) on the delayed posttests, but there was no significant difference between the card group and the other two groups. The questionnaire results indicated that overall, the computer method was viewed more positively, but there was more variation in student perceptions in the PC group than the other two groups. Nakata concluded that both word card study and PC based study are more effective than studying from lists, but stated that due to motivational effects, the PC method is preferable over word cards. Overall, Nakata concluded that the results of the experiment demonstrated that, as learning outcomes should be measured by long-term retention, the PC method led to the highest retention, followed by the Card and List groups. However, as stated in the paper, “we need to treat the results with caution due to the lack of statistically significant difference between the Card group and the other two” (p. 12).

In summary, both deliberate vocabulary study with computers and word cards was more effective than studying from lists. With regard to the efficacy of computers over lists, while the PC treatment group performed slightly better, the differences were not statistically significant when time-on-task was controlled, so the positive claims towards computerized study over word cards are questionable. The limitations of this study include the lack of language proficiency data, and that only one productive L1 to L2 test was used; no information about depth or partial knowledge is available. In addition, whether a period of four days constitutes a delayed posttest is questionable.

In a similar study to Nakata (2008), Oberg (2011) compared the efficacy of paper-based picture cards with a computer-based alternative in terms of vocabulary acquisition and retention. The participants were 71 first-year Japanese university students taken from two intact classes. The two vocabulary study methods investigated were (a) paper word cards containing an L2 word on one side and a representative picture on the other and (b) a teacher-made CALL-based equivalent that presented the picture and a choice of L2 meaning words side-by-side. The students were divided into two groups and studied 10 target items using one of the methods that they had been randomly assigned to. The independent variable was treatment (two levels), and the dependent variable was the score gains between the pre and post multiple-choice cloze tests. Independent samples *t*-test analysis showed that there was no significant difference between the group means on the pretest; $t(69) = .145$, $p = .886$, posttest; $t(69) = 1.130$, $p = .263$, or on the delayed posttest; $t(69) = 1.108$, $p = .272$. In terms of vocabulary gains, the author stated that both study methods resulted in vocabulary gains, but no statistical significance was reported for this; results were discussed only in terms of group means with the CALL group means being approximately one word out of ten higher on both the posttest and delayed posttest. However, a post-treatment questionnaire revealed a slight preference for the CALL method among the participants. Oberg concluded that when the statistical results are taken together with the qualitative data, the CALL method was slightly more effective, and was considered more useful than the picture card method. Oberg concluded that the affordances offered by the CALL method combined with the increased motivational effects due to learner preference for this medium could result in better learning outcomes in natural learning contexts. However, this conclusion is questionable as a higher percentage of students stated that they would want to study vocabulary with the picture cards again rather than with the CALL method. The results are similar to Nakata (2008) in that there was no superiority of a computerized system over a paper-based one in terms of vocabulary acquisition, but learners showed a slight preference for the electronic method. However, the limitations of this study concerned the sample size, the limited number of target items (10), the lack of statistical data regarding the efficacy of each method in terms of vocabulary gains, and that only a single, fill-in-the-blank multiple-choice test was used. Due to these limitations, the results and conclusions must be interpreted cautiously.

Dodigovic (2013) carried out a study with a very similar research focus as the one described in this study. Dodigovic aimed to compare the learning outcomes of studying foreign language vocabulary from teacher-designed electronic flash cards with that of studying from self-designed electronic cards. The main difference between Dodigovic's study and my own was that Dodigovic used digital flash cards, and all information on the cards was written in the L2. The original sample consisted of 102 participants, but unfortunately, due to missing data, the final sample was reduced to 89. All participants were between 17 and 19 years old, were native speakers of Gulf Arabic enrolled in pre-university foundation programs in Qatar and had scores on the TOEFL PBT ranging from 350 to 450. The study involved a within subjects, counterbalanced design with three treatments; (a) learning vocabulary without cards (contrast group), (b) learning vocabulary with teacher-designed electronic cards and (c) learning vocabulary with self-designed electronic cards. Students in the contrast group were presented with the target words through a reading text, and then told to study them however they felt most suitable. The teacher designed cards were presented to the students digitally and time given to study the cards. However, the students created their own cards in the third condition after being given advice on what information to include. The independent variable was the treatment condition, while the dependent variable was the vocabulary learning outcomes, as measured by vocabulary tests. A combination of measurement instruments was used before the treatment, but the only pretest and posttest data came from the VLT. Data were analysed using one-way ANOVA. In terms of reported results, there was a statistically significant difference between treatments on the posttests, $F(2, 86) = 13.67$, $p < 0001$. Post hoc analysis using Tukey's HSD revealed that while there was no significant difference between the contrast group and the teacher-designed card group (q = 0.18, $p < .01$), the student-designed group yielded significantly lower results than either the contrast procedure (q = 4.12, $p < .01$) or the teacher-designed cards (q = 4.31, $p < .01$). Dodigovic stated that the results show that student-designed cards are less effective than teacher-made cards and puts forward the idea that this could be due to affect reasons because students showed a preference towards the teacher-designed cards. While other analyses initially aimed to shed more light on why the mean vocabulary score differences existed, too much missing data for the self-created treatment, for example, time-on-task, and information contained on the cards, prevented this from being carried out. Dodigovic concluded that "students learn best when using either their own favourite strategies or teacher-designed cards" (p. 29). One conclusion is that student-designed cards were less skilfully or fully made than the teacher-designed cards; thus, more training in word card design is needed if student-designed cards are to be more effective.

These results of this study are disheartening for supporters of word cards as neither word card method was better than the contrast method, and self-created cards appeared to be less effective than the non-word card method. However, there are a number of possible reasons why these results need to be read with caution. First, no data showing whether time on task was the same for each treatment are available. It could be that the students in each group spent very different amounts of time during the treatment stage, but this is not known, or at least not reported. Furthermore, the fact that the contrast method involved at least two different study methods; first encountering words in a reading text, then followed by an individually chosen deliberate study technique, intimates that the time spent on the contrast treatment would have been more than on the other two treatments, involved both contextualized and decontextualized study, and could have been viewed more favourably due to the autonomous nature of the study technique chosen. In addition, the lack of sensitive vocabulary pre and posttest instruments, and a lack of statistical data further reduce the ability to judge the results. Overall, while this study addresses an important vocabulary research area, and while many parts of this study were well designed, too many research design and data reporting issues exist for the results to be taken without extreme caution.

In another word card study, Hung (2015) investigated the effectiveness of three types of digital flashcard study techniques carried out within a classroom setting over a period of nine weeks. In addition to examining the vocabulary learning effects of each treatment, Hung also examined students' attitudes towards the different flashcard study methods, focusing on the three aspects of usefulness, ease of use, and intention to use in the future. This research was conducted at a public university in Taiwan, and the sample consisted of 75 Taiwanese EFL students aged between 18 and 20 years old, taken from three intact classes. Hung describes the students being predominantly at an intermediate level of English. As mentioned, the study involved three treatments that were defined by varying task designs of flashcard use during the assigned self-study sessions in class. The three treatments were as follows: (a) a self-practice treatment where participants created and studied from five flashcards on their own, (b) a pair-exchange treatment in which the participants created five flashcards individually and took turns to exchange and study the flashcards with different peers, (c) a group-based treatment where participants first individually made five flashcards before collaboratively deciding on ten useful words to include in their group-based flashcards for later practice and review. The study utilized a pre- posttest design with the independent variable being the treatment condition, and the dependent variables being the vocabulary learning outcomes, as measured by pre and post active recall vocabulary tests, and the data gathered from an attitudinal survey. Data were analyzed using series of one-way ANOVA and post-hoc Tukey tests. The results indicated that while there was no statistically significant difference between the three groups on the pretest, $F(2, 73) = .131$, $p = .878$, there was a statistically significant difference on the posttests, $F(2, 73) = 15.981$, $p < 001$, with students who used digital flashcards collaboratively in groups to learn outperforming those in the other two experimental conditions. In terms of participants' attitudes towards the study methods, analysis showed that participants in the group-based format had significantly more positive attitudes towards the treatment than the other two groups on all subscales of the attitudinal survey: usefulness: $F(2, 73) = 42.767$, $p < 001$, ease of use: $F(2, 73) = 26.549$, $p < 001$, intention to use $F(2, 73) = 54.865$, $p < 001$. Overall, the students learning with the group-based format had a significantly higher level of acceptance about the vocabulary learning task than the other two groups, $F(2, 73) = 118.94$, $p < .001$. Taken together, Hung concluded that not only did the collaborative learning treatments lead to higher vocabulary gains, but were also viewed more positively, and it could be that the positive perceptions resulted in the better learning outcomes. Hung goes on to state that, based on the results, it appears that the higher the level of collaboration, the more positive students' attitudes towards the activities are, and the higher the learning outcomes will be. Overall, Hung concludes that the results of the study are congruent with Nakata (2008); showing that electronic flash cards are an effective method of deliberate vocabulary study, but that this study indicates that their effectiveness can be significantly increased by giving learners the opportunity to work collaboratively with flash cards.

The Hung study is an interesting, well-designed study that offers further empirical support for word cards as a whole, especially digital word cards. The fact that this study also investigates the effects and perceptions of collaborative learning with words cards makes the study quite unique, and the findings are of great interest to the study carried out in experiment two of this study; the learning and motivational effects of the peer-to-peer test system designed by the researcher.

In another study, Nikoopour and Kazemi (2014) carried out an experiment with the aim of evaluating the vocabulary gains and perceptions of advanced learners of English using either digitized or non-digitized flashcards. The participants were 109 male and female Iranian advanced EFL learners aged between 20 and 35 years, who were university students with Persian as their native language. All participants volunteered for the study and are described by the researcher as being motivated to improve their English skills. After taking a pretest, the 109 participants were subdivided into three groups; mobile, online, and paper flashcard groups. Each week for ten weeks, 70 flashcards containing frequent words found on the TOEFL and IELTS tests were made available to study using their assigned study method. Participants took biweekly progress tests aimed at increasing motivation for deliberate vocabulary study. After the ten weeks of treatment, to assess participants' overall vocabulary gains, the participants took a multiple-choice posttest that was identical to the pretest. In addition, participants also answered a survey consisting of 30 Likert scale items to assess their attitude towards the digitized and non-digitized vocabulary delivery. The researchers state that the questionnaire specifically aimed to assess the participants' attitude towards language learning, vocabulary learning, CALL, and mobile learning. The independent variable was study method, defined as three types of flashcard study methods, and the dependent variables were the posttest vocabulary gains, and the attitudinal survey results. Analysis was carried out using a one-way ANOVA, and the results revealed a statistically significant difference between the groups on the vocabulary posttest, $F(2, 106) = 7.787$, $p < .05$. However, to cross-compare the groups and detect where the difference was significant or insignificant, a post hoc comparison was carried out using the Tukey HSD tests. There was a significant difference ($p < .0001$) between (a) the Mobile and Online Groups, and (b) Online and Paper Groups in the results of posttest, however, the difference between Mobile and Paper Group was not statistically significant ($p = .303$). Mean scores indicated that the Mobile Group performed the best on the posttest, followed by the Paper Group, however, as mentioned the mean score differences between these two groups were not statistically significant. However, the Internet group performed most poorly when compared to the other two methods, and both differences were statistically significant. The Mobile group showed a positive mean gain difference of 10.84 words over the Internet Group, and 4.01 over the Paper Group, who in turn had a mean gain difference of 6.82 over the Online Group. In terms of the results of the attitudinal survey, analysis of a one-way ANOVA showed there was a statistically significant difference in the attitudes between the groups, $F(2, 106) = 5.645$, $p = .005$. Post hoc analysis using Tukey HSD tests revealed that there was a significant difference ($p < .05$) in participants' attitude in (a) the Mobile and Online Groups, and also (b) between the Mobile and Paper Groups, but the difference between the Online and Paper Groups was not significant ($p = .954$). The participants in the Mobile Group had more positive attitudes than the other two groups towards vocabulary learning in general, and especially in terms of the treatment method they experienced. The online group had the most negative attitudinal scores of the three groups, although not statistically significant when compared to the paper group.

The researchers conclude that any form of word card delivery system; digital or paper-based, is an efficient and successful way of deliberately learning foreign language vocabulary. However, they go on to state that, given the fact that there was no statistically significant difference between the paper group and the mobile group, but that the electronic group performed the worst, the mode of delivery (electronic or paper-based) is not important, but portability is the key determinant of success. The researchers state that since the flashcards content was the same for all three groups, the superior performance of the Mobile and Paper groups over the Online Group seemed to be due to the portability and high accessibility. With reference to the Online Group, it is stated that “limited Internet access and consequently limited time of study in this group resulted in the lowest mean of scores on the posttest” (p. 1372). In addition, they conclude that as both the mobile and paper-based treatments were viewed more positively by participants, this could also account for some of their success, and offer justification for their inclusion in foreign language curriculums.

Overall, the results of this study offer further support for the use of word cards as a deliberate vocabulary learning methodology. Not only did both mobile and paper-based word cards yield statistically significant learning outcomes, but they were also viewed positively by the participants who used them for ten weeks. This study has many strengths including its experimental design, the fact that it reports both language gains and perceptions related to each method. In addition, as the study was carried out over a ten-week period, and learners studied a large amount of words, the results are more robust than many other word card studies. However, some issues need to be mentioned that could limit the scope or generalizability of the results. First, the vocabulary tests only tested a representative sample of the words studied, so it is unknown if the pattern of results found would apply to all words studied in each condition. Another limitation is that it appears time on task was not controlled for, so it is unclear to what extent study time varied across the treatments. Also, a major difference between this study and many other word card studies is that only L2 information was supplied to the participants; no L1 translations were written onto the cards. Furthermore, while a lot of information was included on the cards, for example, synonyms, antonyms, and sentences, knowledge of this type of information was not tested. The fact that knowledge was only tested with one L2 multiple choice test is another limitation of the study. However, as mentioned, many other positive research design aspects make this a strong study, and the results definitely add needed information to the word card research literature.

Another study investigating the comparative efficacy of paper-based and digital word cards was carried out by Emrah and Omur (2010). The researchers compared mobile phones and traditional paper flashcards for deliberate vocabulary study. 60 university students, who were studying in the Black Sea region of Turkey, were divided into two groups. One was the experimental group who used mobile phones, and the other was the comparison group who used paper flashcards, a traditional vocabulary acquisition technique for vocabulary studying. The vocabulary items were selected from the ECTACO Flash Card application, and participants were required to study the vocabulary for six weeks. Data were analyzed independent and dependent samples t-tests. The mean scores and statistical significance for both mobile phones (M = - 13.29, $t(28) = -7.6$, $p <$

.05), and traditional paper flashcards (M = -8.62, $t(28) = -5.7$, $p < .05$) showed that both methods were effective as scores were statistically higher than on the pretest. Therefore, the first result of this study was that both methods were effective for learning vocabulary. However, further analysis produced a significant effect in favour of the mobile flash card system. The mean gain score for learners in the experimental group (M = 13.79) were significantly higher than the mean gain score for the comparison group (M = 8.62), $t(56) = 2.191$, $p < .05$. The researchers also reported that analysis of qualitative data, gained through semi-structured interviews with eights participants, showed that mobile phones were the preferred method of study. Overall, Emrah and Omur (2010) concluded that using mobiles phones was more effective than using paper word cards in terms of both vocabulary gains and motivation, and therefore, recommend using mobile phones for English vocabulary learning.

While this is an interesting study that lends support for both traditional and digital word cards, there are some research design issues that need to be taken into consideration, especially in terms of the comparative data. First, the purpose of the study, and the researchers' expectations from the students during the experiment were explained before the study, something that might have affected the participants' actions. This could be a large problem if the researchers clear pre-study bias towards the mobile study method was noticeable to the students. Second, it is not clear what form the paper word cards took, for example it is not mentioned if the L1 and L2 word forms were written on opposite sides of the cards. If this was not done, given that it is the principle design requirement of word cards, this cannot be considered a word card study. All that is said about the word card treatment design is that the comparison group "was given the vocabulary lists on paper and they memorized them through self-study" (p. 4).

In another word card study, although one limited to the use of electronic word cards, Altiner (2011) investigated the learning outcomes of studying vocabulary with a computer-based flashcard system (Anki). The participants were 13 university students in Iowa State University. The participants were Mandarin Chinese native speakers and had studied English for several years. She used the Academic Word List and asked the participants to study 20 new target words for 10 minutes every day for three weeks. Students were given an active recall pre and posttest. Data were analyzed using paired samples *t*-tests, and qualitative data gained from a survey were analyzed and tabulated. In terms of results, the Anki computer flashcard program was deemed to be effective as mean scores on the pretest were significantly lower than those of the posttest; (M = -4.3, SD = 3.7); $t(11)$ = -3.9, p = .002. Altiner stated that the results show that studying from computer-based flashcards is effective for university students learning a foreign language. Altiner concluded that the process of learning academic vocabulary with Anki helped college level ESL students to improve their academic vocabulary and changed their negative perceptions of deliberate vocabulary study. This is an interesting study, however, given that the number of participants was very small, the results need to be taken with caution.

In the most recent word card study, and one that is very much related to the studies contained in this study, Reynolds and Shih (2019) investigated the learning outcomes of using self-constructed word cards. In addition, they used survey data to assess participants feelings about using self-constructed word cards for deliberate vocabulary study, and to assess students' vocabulary self-selection strategies. The participants were 100 Taiwanese English as a Foreign Language (EFL) 10th graders. The study adopted a pretest-posttest design and used the Vocabulary Levels test (VLT) and the Specific Vocabulary Test (SVT), a teacher-designed 50-item test, were used as measures of vocabulary knowledge. The intervention group ($N = 50$) were given 18 weeks of classroom English reading skills instruction in addition to homework requiring construction and review of 100 vocabulary word cards using unknown words encountered during reading activities. The word cards included the target L2 word, their L1 meaning, and L2 synonym, and a representative picture. In contrast, the control group ($N = 50$) received only the skills-based reading instruction for the same 18-week period. Reynolds and Shih found significant differences in vocabulary acquisition (SVT) between the two groups after the 18-week treatment stage, with the word card group performing better, $t = 5.51$, $p < .001$, $df = 98$, 95% CI = -27.48, -12.92, but no difference in vocabulary size (VLT). In addition, survey responses revealed that while participants viewed word cards as effective for learning the meaning of unknown L2 words, they had negative perceptions towards using self-constructed word cards for deliberate vocabulary learning. However, as mentioned, this study used word cards that involved learners drawing pictures on the cards, and it was this aspect that participants seemed to have the most negative feelings towards. In addition, they found that learners need guidance in selecting target words from reading texts as many students selected words that were not appropriate for their abilities. Reynolds and Shih concluded that (a) the self-construction and review of word cards is an effective strategy for increasing EFL learners' vocabulary knowledge, (b) the vocabulary self-selection strategy needs to be supervised or guided if it is to be effective, and (c) if learners have the benefits of self-created words explained to them, it might increase motivation.

This is a very interesting and clearly relevant study, and the main strengths of this study are that it used a longitudinal design and collected survey data in addition to learning outcome data. However, the researchers did not use such accurate measures of vocabulary acquisition (the tests were not based on words the learners had actually chosen and used to make and study form the word cards), and the survey included only one open question. Also, very little detail is given about the qualitative data gained from this question. In addition, given that learners were required to draw pictures, the word cards used by Reynolds and Shih are very different to those most teachers and researchers have in mind when thinking of word cards (Nakata, 2020, pp. 304, 307) and as such, they are also very different to those used in this study.

Summary of the Studies on Word Cards and Decontextualized Study Methods

In summary, the studies reviewed above provide empirical evidence in support of word cards and decontextualized activities for vocabulary acquisition. The results of the various word card studies reviewed above are presented in Table 4. In addition, as some of the studies deal with the broader issue of decontextualized learning, therefore, Table 4 also presents the results of the relevant studies reviewed above along with others discussed in the following sections.

Table 4. *Results of Studies on Word Card and Decontextualized Study Methods*

Word Cards	
Word card study results in significant vocabulary knowledge gains	Altiner (2011); Dodigovic (2013); Emrah & Omur (2010); Hung (2015); Komachali & Khodareza (2012); Kuo & Ho (2012); Nakata (2008); Nikoopour and Kazemi (2014); Reynolds & Shih (2019), Waring (1997)
Word cards are a time efficient method of deliberate vocabulary study	Komachali & Khodareza (2012); Kuo & Ho (2012); Mondria & Mondria-de Vries (1994); Nakata (2008); Waring (1997)
Word cards help students learn both receptive and productive vocabulary knowledge	Komachali & Khodareza (2012); Kuo & Ho (2012); Waring (1997)
Word card learning is superior to list learning	Kuo & Ho (2012); Mondria & Mondria-de Vries (1994); Nakata (2008)
Learners show positive perceptions towards word card use	Altiner (2011); Dodigovic (2013); Hung (2015); Kuo & Ho (2012); Nikoopour and Kazemi (2014);
Decontextualized Learning	
Learning words with little or no context is effective for word recall	Altiner (2011); Dodigovic (2013); Hung (2015); Komachali & Khodareza (2012); Kuo & Ho (2012); Laufer & Rozovski-Roitblat (2014); Laufer & Shmueli (1997); Mondria & Mondria-de Vries (1994); Mondria & Wiersma (2004); Nakata (2008); Emrah & Omur (2010); Waring (1997)
Vocabulary focused activities yield superior amounts of learning than incidental activities such as reading	Eckerth & Tavakoli (2012); Laufer & Rozovski-Roitblat (2014); Webb (2005)

The various studies have shown that word card learning results in significant vocabulary knowledge gains in relatively short amounts of time, is superior to learning from word lists or from reading passages and is viewed positively by learners. While some studies suggest that computerized methods might be better (Emrah & Omur, 2010), not enough significance data are provided to come to this conclusion. Therefore, the results of the experiments discussed above offer research-based support for the belief that language teachers and learners are justified in choosing word cards as a deliberate study method (Emrah & Omur, 2010; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Reynolds & Shih, 2019). However, as discussed below, regardless of the considerable support for word cards, there are still some very important issues relating to word cards that have not been empirically investigated or even questioned.

Productive (Writing) Versus Receptive (Reading) Learning Techniques

Hulstijn and Trompetter (1998, pp. 191–200) compared the incidental vocabulary learning outcomes of dictionary use during a computerized reading task with that of a computerized productive writing task. The participants were 110 native Dutch speaking high school learners taken from four classes in four high schools. Two intact classes (50 learners) were assigned to the reading condition and two classes (60 learners) were assigned to the productive (writing) condition. For the treatment, the participants in the reading group carried out an incidental vocabulary learning activity using a computer-assisted weather report reading task, followed by comprehension questions, while the participants in the writing group carried out a productive task, in which they were provided with an L1 example glossed in the L2 and instructed to write their own weather report in the L2 incorporating the target vocabulary items. In both treatments, the participants were instructed to look up unknown words in a dictionary. An innovative part of this study was that individualized tests were created based on the words that the participants had looked up. The independent variable was treatment, the dependent variables were the number of words looked up in the dictionary, and the scores on receptive and productive post-treatment vocabulary test of these items. The data were analyzed using multiple *t*-tests, the type of *t*-tests was not mentioned.

Results were that out of ten words in each condition, the mean vocabulary retention score for the reading group was 3.8, and that of the productive group was 4.6. However, there was no statistically significant difference between the two groups, $t(108) = 1.47$, $p = .55$. The researchers stated that contrary to their hypothesis, the productive task was no more effective than the reading task as no statistical difference was found between the group mean vocabulary scores on the two task types. However, the researchers then stated that the higher mean scores for the writing group offer modest support for the view that when processing new lexical information, writing activities lead to more elaboration and therefore more retention. However, this opinion can only be accepted if (a) the lack of statistical significance is overlooked, and (b) time-on-task is not considered an important factor. This is because the writing group took significantly longer (31.6 minutes) than the reading group (21.5 minutes), and this time difference was statistically significant, $t(118) = 6.56$, $p < .001$.

This study supports both receptive and productive vocabulary learning approaches, because while the gains were not overly impressive with approximately four words learned in 20 to 30 minutes, learning did occur. However, this study has a number of issues that must be considered when interpreting the results. First of all, it is questionable whether or not two incidental learning tasks were compared because the writing group "were encouraged to consult the computer in order to look up L2 weather terms" (p. 194). While this task does fall into Hulstijn's definition of incidental learning, it could be argued that there is an intentional vocabulary learning element. In addition, a major weakness of the study is that, due to the absence of pretesting, there is no way to know whether the participants really did not know the words they looked up. For example, they might have had some knowledge of the words and were simply confirming their understanding when they looked up the meanings in the dictionary. In addition, time-on-task was not controlled, therefore it is difficult to make direct comparisons of the efficacy of the two treatments.

Overall, this study supports both reading and writing tasks for incidental vocabulary learning, but due to methodological flaws, it is unclear how much learning took place as a result of either learning conditions, and whether or not they accurately represent incidental learning.

In another important deliberate vocabulary study, Webb (2005), carried out two experiments to investigate the effects of receptive and productive vocabulary learning on 10 types of word knowledge. In this study, receptive learning again refers to reading activities in which learners see word presented in sentences, while productive learning refers to the writing of sentences using target words. For both experiments, Webb used nonsense words that were matched with the meanings of low-frequency L2 words as the target words because this guaranteed no prior knowledge of target words, and therefore eliminated the need for pretests. Five aspects of vocabulary knowledge—orthography, syntax, association, grammatical functions, and meaning and form—were measured using a variety of receptive and productive tests, with 10 tests being used overall. The scores on the 10 posttests represented the dependent variables for both experiments. Both experiments were conducted within one 90-minute class period in the usual teaching context.

In the first experiment, 66 Japanese native speakers from one first-year EFL class learned 10 target words either from three glossed sentences, or in a sentence production task. Therefore, the learners were divided into two experimental groups, a receptive group and productive group, and this represented the independent variable. Learners in the both treatments were exposed to each target word and their L1 meaning, however, learners in the receptive group were also shown three glossed sentences with target items underlined and written in bold.

Learners in the receptive treatment were simply told to learn the words, whereas those in the productive treatment group were told to (a) write each word in an English sentence and (b) learn each word. In the first experiment, each group was given 12 minutes to complete the treatments before taking the tests, however, based on pilot testing, Webb stated that this gave just enough time to complete the productive task, but more than the amount of time needed to simply read through the receptive materials. Analysis was carried out using a MANOVA using the scores on the 10 dependent measures and the independent variable of type of learning task. The results of the first experiment showed that while both treatments resulted in increased vocabulary knowledge, when the same amount of time was spent on both tasks, the reading task produced superior word knowledge overall, $F(10, 55) = 5.07$, $p < .001$. Learners who completed the receptive (reading) task had higher scores than the productive group on all of the dependent measures, but the differences were only significant for five of the dependent measures (see Webb, 2005, pp. 43-44 for details of the statistical data).

The second experiment used the same materials and tests, but the learners were given only enough time to complete each task. The aim of the second experiment was to examine if the results from Experiment 1 held true when learners only spent the time needed to complete the tasks. Forty-nine first-year Japanese university learners learned 10 target words under each treatment condition; they were instructed to begin the second task as soon as they had completed the first. A repeated measures MANOVA was performed with the independent variable again being the type of learning task (receptive or productive), and the 10 dependent measures were used to analyze the within-subjects factor. Conversely to the results of Experiment 1, the results showed that the writing task resulted in significantly higher scores than the reading task, $F(10, 39) = 2.179$, $p < .05$. The positive mean differences of the productive group over the receptive group were significant for all dependent measures (see Webb, 2005, p. 47 for details of the statistical data). Webb stated that the second experiment showed that if time-on-task is not a concern, and it is accepted that the writing task takes longer to complete, then the writing task is more effective, a finding also reported in the more recent Webb and Piasecki (2018) study. Webb concluded that the second experiment might represent more authentic learning as learners will not spend more time completing a task than needed, and as such, a stronger argument can be made in favor of productive vocabulary learning tasks over receptive tasks in natural teaching contexts. Webb stated that productive tasks are especially important because it is well known that learners' receptive vocabulary knowledge is less than their productive knowledge, and therefore, productive activities that can reduce this gap are needed in language curriculums. Taken together, these two studies offer further support for the use of both receptive and productive deliberate vocabulary study tasks. In addition, the results show that both modes result in learning in the other mode; receptive learning can result in productive knowledge and vice versa. The results also show that even though productive activities take more time, they provide significant learning opportunities and could be more effective than receptive tasks alone.

Barcroft (2004) investigated the comparative effects of learning words only receptively (visually) with that of writing words in sentences after the initial visual presentation. There were two experiments in this study, the first of which did not consider equality of time on task and number of repetitions as a necessity. However, the second experiment controlled for time on task and number of repetitions. Each experiment and the results are described separately below.

For the first experiment, the participants were 44 second-semester students from four different course sections at an American university. All participants spoke English as their first language and were studying Spanish as an L2. The experiment used a within-subjects design with all participants studying 12 words only receptively, and another 12 words through a combination of initial receptive exposure followed by productive sentence writing task for each word. The researcher utilized a counterbalanced design to limit both word and order effects. In the no sentence writing condition, learners were exposed to the target words for a total of 24 seconds over four repetitions, while the sentence writing group was shown the word for 48 seconds with only one exposure. In the no sentence group learners were simply told to try to learn the L2 words presented along with a representative picture on a television screen. However, in the sentence writing group, the participants were told to write a sentence using the word after the initial presentation. The experiment used a pretest-posttest design with a pretest immediately before the treatment and one posttest directly after the treatment, another test two days later, and a final test one week after the treatment. The same test, which was a productive vocabulary test, was used for all test times. The test performance was scored in two different ways; a syllable scoring method where participants received points for each correct target item syllable written, and a word scoring system where one points was given only if the whole word was written correctly in the tests. The within-subject independent variables were condition and time, and test scores acted as the dependent variables. The results of repeated measures ANOVA based on the syllable scoring revealed a significant main effect for condition, $F(1, 43) = 86.01$, $p < .001$, $\eta^2 = .667$, and for time, $F(2, 86) = 40.56$, $p < .001$, $\eta^{2 =}$ .485. Post-hoc analyses indicated that the means for the no sentence writing condition were considerably higher than the means for the sentence writing condition. Overall means were 9.64 for no sentence writing and 4.40 for sentence writing, which is a difference of 119%, and the negative effect of sentence writing was maintained over time. The results of repeated measures ANOVAs based on the word scoring followed a similar pattern and again revealed a significant main effect for condition, $F(1, 43) = 84.97$, $p < .001$, $\eta^2 = .664$, and for time, $F(2, 86) = 9.88$, $p < .001$, $\eta^{2 =}$ .187. Analysis indicated that the means for the no sentence writing condition were again considerably higher than the means for the sentence writing condition. Overall means were 2.75 for no sentence writing and 1.25 for sentence writing, which is a difference of 113%. Again, the negative effect of sentence writing was maintained over time with this group performing worse on all posttests.

For the second experiment, the participants were 10 students from two different course sections at a different American university. Again, all participants spoke English as their first language and were studying Spanish as an L2. The experimental design including words, treatments and tests was the same as that described in experiment 1, with the only difference being that the amount of exposure time and number of repetitions was controlled for. Therefore, in experiment 2, each word in each group of 12 words was presented only once for 24 seconds. As in the first experiment, the within-subject independent variables were condition and time, and test scores acted as the dependent variables. The results of repeated measures ANOVA based on the syllable scoring revealed a significant main effect for condition, $F(1, 9) = 12.18$, $p = .007$, $\eta^2 = .575$, but not for time, $F(1, 9) = 4.13$, $p = .073$, $\eta^{2\,=}$ .315. Analysis indicated that as with experiment 1, the means for the no sentence writing condition were considerably higher than the means for the sentence writing condition. Overall means were 10.3 for no sentence writing and 4.95 for sentence writing, which is a difference of 108%, and these results again followed a similar pattern on each posttest. The results of repeated measures ANOVA based on the word scoring followed a similar pattern in terms of overall means, but no significant main effect for condition or time was found. Overall means were 2.5 for no sentence writing and 1.6 for sentence writing. The lack of statistical significance was attributed to the small sample size and the lack of precision of whole word scoring as a measure of degree of word form knowledge. However, Barcroft points out that while not statistically significant, the means for the no sentence group were consistently higher and therefore in line with all other significant findings.

Overall, Barcroft concludes that the results show strong evidence that sentence writing, when no context is provided, has a negative effect on deliberate word learning. He suggests that this is due to the fact that writing sentences uses too many cognitive processing skills that detract from learning the words. He states that the increased semantic processing associated with writing words in sentences clearly reduces form-oriented learning. He proposes that the simpler receptive learning tasks allow learners to use more cognitive processing to make word-meaning connections and focus on the more formal properties of the target words. He also proposes that while semantic elaboration tasks such as writing sentences, can facilitate memory of known words, they might hinder memory for new word forms.

The results of this study are of interest because when making word cards, many learners write example sentences on their cards, something that might not be beneficial according to the results of this study. However, as other research has shown the benefit of presenting words in context (Hulstijn & Trompetter, 1998, pp. 191–200; Webb, 2005), when taking the results of these studies together with those of Barcroft (2004) above, it could be concluded that if learners wish to have example sentences on their word cards, pre-made cards might be more effective. A limitation of the Barcroft study is that the treatment does not represent the type of learning condition that students would carry out in real-life sentences. However, it is a legitimate learning method that offered a high degree of control over confounding variables.

In two similar experiments, Barcroft (2006) compared the effects of writing L2 target words and vocabulary learning. While Barcroft (2004) focused on the effect of sentence writing, this study compared learning words only receptively versus study both receptively and productively; first studying receptively and then writing only the words. The sample for the first experiment included 20 English-speaking L2 Spanish students from two second-semester Spanish classes at a large public university in the Midwestern United States. Experiment 2 included 46 English-speaking L2 Spanish students in their second semester at a private university in the same area of the United States.

Both experiments used a within-subjects design where students had to learn 12 words by simply being exposed to them twice and learn another 12 words by writing them down after the initial exposure. All words were presented in the form of picture and L2 meaning flash cards. The treatment (no writing, word writing) and time (immediate, 2 days later) were within-subject independent variables, and the dependent variable was vocabulary scores on the productive vocabulary tests. Results were analyzed using one-way analysis of variance (ANOVA) with the Alpha being set at .05 for statistical significance. For experiment 1, results of the ANOVA revealed significant main effects for condition, $F(1, 15) = 9.06$, $p = .009$, and for time, $F(1, 15) = 31.78$, $p < .001$. The results indicated that word writing had a negative effect on productive L2 vocabulary learning on both the immediate and delayed posttests. Results followed the same pattern for experiment two, with the word writing condition again producing lower test score means on both posttests: Condition = $F(1, 45) = 10.43$, $p = .002$, and Time = $F(1,45) = 99.80$, $p < .001$.

Barcroft stated that the findings of both experiments indicate that copying target words negatively affects productive L2 vocabulary learning. He concluded that copying words without any context or additional meaning information added an additional cognitive burden that resulted in lower levels of learning of target items. He proposes that the cognitive requirements of copying a word might take up processing resources learners could have used to encode word forms and developing form–meaning connections. He proposes that in order to learn a new word, learners must allocate sufficient processing resources toward encoding the target word forms, isolating appropriate semantic representations, and mapping word forms onto the semantic representations, but the ability to do this is restricted during forced output vocabulary tasks. Overall, this was a well-designed and interesting study, the results of which call into question the practice of having learners make their own word cards. If Barcroft's results hold true, it might intimate that learners would be better served by studying from pre-made cards instead of having to write the information on the cards themselves.

In another Barcroft study concerned with the effect of writing (copying) words during the learning stage, Barcroft (2007b) investigated the learning effect of writing full words, writing word fragments, and not writing the words at all. Again, participants were English L1 students studying Spanish as an L2 at an American university. From an initial sample of 48, only 21 students completed the treatment and all posttests. After taking a written receptive (L2-L1) pretest, participants attempted to learn eight words in each condition, with each target word first being presented alongside a representative picture for six seconds. Then, after the initial exposure, all words were displayed again for a further six seconds in the same order. For the no writing condition, learners were asked to do their best to learn the words, in the writing condition they were asked to copy the full word, and the fragment condition required them to copy the underlined part of the target word. Immediately after the treatment, learners took an immediate written productive (Picture to L2) and written receptive (Picture to L1) posttest, and then took the same tests again two days later but with the order of items changed. The experiment used a counterbalanced design to reduce word and order effect; participants were divided into three groups of seven and each group carried out the three treatments in a different order and learned different words under each condition. The results were analyzed using repeated measure ANOVA with condition (no word writing, word writing, fragment writing), recall type (picture-to-Spanish, Spanish-to-English), and time (immediate, two days later) as within-subject independent variables, and presentation order as a between-subjects independent variable. The dependent variable was the mean scores on the various tests. None of the participants were able to translate any of the target items on the pretest so no scoring protocol was needed, however, posttest scores were analyzed using two different scoring protocols. First, a lexical production scoring protocol (LPSP-written) that gave partial credit for partially correct answers was used. In addition, in order to compare the results with those of a similar study (Thomas & Dieter, 1987), the scoring methods described in their study were also used.

The main results of the Barcroft study were that the repeated measures ANOVA revealed a significant main effect for condition, $F(2, 18) = 12.84$, $p < .001$, $\eta^2 = .42$, and a significant main effect for recall type, $F(1, 18) = 20.34$, $p < .001$, $\eta^2 = .53$. In addition, a significant Order x Condition interaction was also found, $F(4, 36) = 5.56$, $p = .001$, $\eta^2 = .38$, and this was due to substantially lower scores for word writing compared to the other conditions in Order 2. Pairwise comparisons revealed that scores in the no-writing condition were significantly higher than in the word-writing condition, $p = .010$, and for fragment writing, $p < .001$. In addition, scores were significantly higher for word writing than for fragment writing, $p = .013$. The results based on the Thomas and Dieter (1987) scoring method were very similar, with one scoring method showing a significant main effect for condition, $F(2, 36) = 6.14$, $p = .005$, $\eta^2 = .25$, and the other method revealing a significant main effect for condition, F(2, 36) = 5.24, $p = .010$, $\eta^2 = .22$, for time, F(1, 18) = 4.59, p = .046, $\eta^2 = .20$, and a significant Condition x Order interaction, $F(4, 36) = 3.46$, $p = .017$, $\eta^2 = .28$. Under these scoring methods, scores on the no writing condition were again significantly higher than for the writing and fragment wiring conditions.

Overall, Barcroft stated that the combined findings of the various analyses were that vocabulary learning was higher for the no writing condition than for the word writing and fragment writing condition, and higher for word writing than for fragment writing. Barcroft stated that the results provide further evidence that writing words without any context has a negative effect on the learning of these words, and that these finding are consistent with the predictions of the resource depletion for output hypothesis. In a similar conclusion to those drawn in his other studies on word writing (2006), and sentence writing (2004), Barcroft concluded that "copying words can decrease vocabulary learning by exhausting processing resources that otherwise could be allocated toward learning new word forms and making connections between new word forms and their meanings" (p. 723). The above results are directly related to Experiment 1 described in this study because if they are true, they suggest that studying from premade word cards might be more effective than making the cards oneself.

Next, Behzadi and Haji Pour Nezhad (2014) investigated what they referred to as the popular belief that words learned through both receptive and productive tasks are better retained receptively than words that are only learned receptively. The participant sample comprised of 75 upper-intermediate adult EFL Farsi native speakers who were randomly divided into three treatment groups of 25. The independent treatment variable was study method with three levels, and the dependent variable was the scores on the pre-, post- and delayed multiple-choice productive and receptive posttests. One group received two receptive reading tasks, the second group learned words productively, and the final group learned the words through a combination of both task types. No information is given about when the pretest was administered, but the posttest was administered directly after the learning session, and the delayed posttest three weeks later. The three tests were identical except the order of the items and distractors were changed each time.

In the receptive group, participants encountered each of the 20 target words along with their first language (L1) meaning and two example sentences (receptive). They were then given a receptive task that presented each of the target words in three new sentences and were then read silently. In the productive group, the participants encountered the words in the same sentences as the receptive task, however, "the node word was provided alone and the students were required to find the appropriate collocate from the list of words provided and read the words aloud as they filled in the blanks" (p. 15). For the combination group, they first received the 20 target words along with their first language (L1) meaning and two example sentences, and then had to carry out the productive task in the same way as the productive group.

In terms of results, the authors claimed that all three treatment groups made significant gains on the immediate and delayed posttests; however, no statistical data are provided. The results are simply reported in terms of descriptive statistics. The mean gains on the immediate posttests were approximately 18 words out of 20, with all groups performing similarly. The delayed posttest means were between 14 and 18, with the receptive group performing the best on the receptive test, but worst on the productive test. Between-subjects MANOVA results were reported, however, only significance values were reported. The only significant difference between the groups was on the delayed productive test, with the receptive + productive task group outperforming the receptive and productive task groups, $p < .001$. The researchers concluded that learning words both receptively and productively leads to a similar level of receptive retention as learning just receptively, and therefore offers little advantage, especially as this takes more time. However, they stated that learning words through activities that combine receptive and productive study methods are most effective for improving productive knowledge. They went on to state, like Webb (2005), that as most learners' main aim is to be able to communicate productively, activities such as the combination method used in this study should be adopted within language curriculums.

However, a number of research design and reporting issues mean that the results of this experiment need to be interpreted with caution. First, no information is given regarding whether time-on-task was controlled or not. Also, there is little information provided in the paper about the multiple-choice test items, and no information is given about the design of the productive test. Finally, the statistical reporting is limited with regard to the immediate posttest results; only mean and standard deviation data are provided, and the lack of MANOVA data makes traditional reporting of these statistics impossible. Taking all of the above into consideration, while the results of the study support both receptive and productive tasks and are in line with findings of other studies such as Webb (2005) regarding the efficacy of productive tasks over receptive ones, this study has a number of serious flaws that limit its reliability. Therefore, results must be interpreted with caution.

In summary, the various studies above have produced somewhat mixed results regarding the efficacy of receptive reading and productive writing tasks, but overall, they indicate that receptive tasks are more beneficial when we consider time-on-task. While Hulstijn and Trompetter (1998, pp. 191–200) showed that incidental writing tasks produced slightly better learning outcomes than receptive tasks, no significant differences were found. In addition, the writing tasks took longer, so it is likely that the reading activity would have produced similar if not better word learning outcomes if carried out for the same amount of time. Similarly, Webb (2005) found that productive tasks only produced significantly higher results because they take longer. When learners were engaged in the tasks for the same amount of time, reading produced significantly higher word learning outcomes on a variety of sensitive measures. The results of these two studies do offer support for writing tasks, but only as they force the learners to spend more time-on-task. However, based on the results of a series of experiments, Barcoft (2004, 2006, 2007b) states that writing tasks result in lower learning outcomes than comparable reading tasks. However, Barcroft's experiments focused on decontextualized writing where no context was given, so we do not know if his results would hold true when context is provided; when learners copy sentences given to them. In addition, the treatments used in the Barcroft study do not represent any real-life learning methods; no learners would attempt to study words under his experimental conditions. Finally, Behzadi and Haji Pour Nezhad (2014) concluded that learning words both receptively and productively results in better productive retention but offers no benefit over receptive-only study in terms of receptive retention.

Overall, while the results are mixed, all of the studies except those carried out by Barcroft offer some support for productive writing activities, if for no other reason than writing activities result in more time-on-task. The results of the Hulstijn and Trompetter (1998), Webb (2005), and Behzadi and Haji Pour Nezhad (2014) can be summarized by Webb (2005) when he states that if it is accepted that writing task takes longer to complete, then writing tasks are more effective. In addition, as Behzadi and Haji Pour Nezhad (2014) note, as most learners' main aim is to be able to communicate productively, productive activities such as those reported above should be adopted within language curriculums. However, given the mixed results of these studies, more research is needed to show the different learning outcomes attributable to the two approaches. Research is especially needed that looks at this issue with regard to deliberate learning tasks, especially those that use real-life study methods as the treatment, unlike those used in the Barcroft studies.

Receptive (L2-L1) and Productive (L1-L2) Vocabulary Learning

When learning new words, learners often encounter the foreign language (L2) word during reading or listening tasks, and then look up its meaning in the native language (L1). This process is referred to as receptive learning. On the other hand, if learners realize that they do not know the L2 form of an L1 word needed in writing or speaking tasks and therefore look it up in the dictionary, this is called productive learning. One of the advantages of word cards is that, depending on learner preferences, target words can be studied either receptively or productively. However, few researchers have investigated whether one learning method is superior to the other. While not central to the experiments in this study, qualitative data collection methods will investigate whether learners have a preference for either method. Therefore, the findings of such studies investigating this issue are of interest and some key studies are discussed below.

First, the Waring (1997) study described earlier investigated the learning and retention outcomes of learning from word cards receptively and productively. The 70 female Japanese learners were supplied with two sets of 15 words cards, each containing the L1 and L2 form of a low-frequency word on different sides. The participants then learned the word sets either receptively (L2-L1) or productively (L1-L2). Using repeated measures analysis of variance (MANOVA), Waring found a significant main effect for type of learning, with receptive learning yielding higher test scores ($F(1, 56) = 115.2$, $p = .0001$). Waring concluded that receptive learning might be superior to productive learning as it yielded significantly better results, it is easier than productive learning in that it takes less time to gain knowledge of target items, and receptive knowledge seems to suffer less decay than productive knowledge.

Mondria and Wiersma (2004, pp. 79–100) examined the effect of receptive (L2-L1) and productive (L1-L2) deliberate vocabulary study on receptive and productive word knowledge. The study was primarily carried out to test what the authors state is a common belief that words studied both receptively and productively are better retained receptively than words studied receptively only. The participants were 198 Dutch pupils aged between 14 and 16 years who were studying French as a foreign language. There were three levels of the independent treatment variable: receptive, productive, and receptive + productive. The dependent variables were scores on the immediate and delayed receptive and productive posttests. The learners were required to study 16 target French words in 15 minutes either receptively (Dutch to French) or productively (French to Dutch) from a list. All learners were tested on the 16 target words in both modes; they were required to translate all words from French to Dutch, and Dutch to French.

The first finding was that there was no significant difference between the receptive + productive group and the receptive only group on receptive retention, $t(65) = 0.36$, $p = .36$ (one-tailed test). However, the productive + receptive group performed significantly better than the receptive only group, $t(65) = 1.98$, $p = .03$ (one-tailed test). With regard to productive retention results: the receptive + productive group performed significantly better than the receptive only group, $t(64) = 3.03$, $p < .01$ (one-tailed test), but the productive + receptive group did not perform significantly better than the productive only group, $t(64) = .32$, $p = .75$ (one-tailed test). Finally, while receptive learning did not produce better receptive knowledge than productive learning, $t(64) = 1.52$, $p =.07$. (one-tailed test), productive learning did lead to significantly higher productive gains than receptive learning, $t(62) = 3.2$, $p < .01$ (one-tailed test). In summary, there were mixed results regarding the effect of the combination method efficacy, but overall, productive learning offered stronger benefits. Overall, the researchers stated that the results indicated that (a) all of the study modes resulted in significant gains in both receptive and productive knowledge, (b) but surprisingly, studying words both receptively and productively did not offer any significant benefits, especially when time-on-task is taken into consideration. The researchers concluded that either mode is effective for producing receptive and productive knowledge, but that as receptive learning is easier and takes less time, it might be the optimal mode for learners given the vast number of words that need to be learned to gain high levels of foreign language competency.

In summary, the results showed that studying words from the L1 to L2, or L2 to L1 are both equally effective, but it is not necessary for learners to do both. However, in a natural learning context, most learners first study receptively as they either meet the L2 form of a word in a text (incidental learning) or word list (deliberate), and then translate it into the L1. Then, if they choose to study these words deliberately, for example, using word cards or notebooks, it is most likely that they will do a combination of both receptive and productive learning, both of which were shown to be effective in this study. Therefore, the results of this study offer further support for deliberate vocabulary study techniques that (a) provide learners with autonomy in how to study (receptively or productively), but (b) offer the opportunity for both.

Laufer and Shmueli (1997) examined the relationship between short- and long-term memorization of new words, and teaching techniques involving different modes of vocabulary presentation and L1 and L2 vocabulary glossing. The participants were 128 high school pupils studying in the highest-level English classes whose L1 was Hebrew. The 128 students were divided into four treatment groups and one comparison group. The four treatments were words presented in isolation, in minimal context (in one meaningful sentence), in a text-context (reading passage with 20 words glossed in the margin, 10 L1 and 10 L2 glosses), and in an elaborated text context (in the original text supplemented by clarifying phrases and sentences). In each mode of presentation, 10 words were translated into learners' L1 and 10 were explained in English. The comparison group was asked to learn the words for a quiz by themselves. The independent variable was group, and the dependent variables were the scores on immediate and delayed L2 multiple-choice posttests (short- and long-term retention). A pretest was not used, but the researchers stated that pilot testing ensured the low-frequency target words would be most likely unknown to the participants. The data were analyzed using a series of one-way ANOVAs and matched *t*-test, and retention scores were compared by mode of presentation, language of glossing, and the interaction between the two. The results of most interest are the long-term retention effects of the treatments.

In terms of the mode of vocabulary presentation, the results of a one-way ANOVA indicated the most effective method for long-term retention was the sentence method, $F(4, 110) = 13.34$, $p = .0001$, followed by lists. With regard to the language of presentation, matched *t*-test results indicated that words glossed in L1 were better retained in the long-term than those glossed in the L2 in all presentation methods, $t(127) = 6.37$, $p = .001$. In terms of interaction effects between method and language of presentation, the ANOVA results showed that there was no interaction between method and language of glossing on short-term retention, $F(3, 94) = 1.54$, $p = .20$, or on long-term retention, $F(3, 87) = 0.14$, $p = .94$. Lack of interaction results show that none of the methods favor a specific language of presentation. Overall, the results were that in terms of context effect, words presented in sentences and lists were remembered better than words presented in text and elaborated text. In the researchers' words, the results "underscore the importance of attending to newly learnt vocabulary and relating it to the first language" (p. 1). The researchers also suggested that mental elaboration, something credited with affecting retention, might not occur when learners encounter words in texts; however, bilingual lists can result in this kind of elaboration. The important findings of this study are (a) L1 translations are beneficial to word learning because they provide support for receptive learning (L2-L1), and (b) words presented with no or little context are better remembered and understood than those presented in reading passages; learning from word pairs and glossed sentences was superior to learning from glossed or elaborated texts. Overall, the findings support the use of deliberate vocabulary study, especially techniques involving the use of L1 translations and example sentences.

As a summary of the results, while all three studies show the benefit of both receptive and productive study, the results especially offer support for receptive study, at least in the initial word learning stage. Waring (1997) found that receptive learning was easier as it leads to higher scores even though it took less time. Mondria and Wiersma (2004, pp. 94-95) concluded that both approaches resulted in significant gains in both receptive and productive knowledge but studying words both receptively and productively did not offer any significant benefits when time-on-task was taken into consideration. The researchers concluded that either mode is effective for producing receptive and productive knowledge, but their overall conclusion was in line with Waring's, as receptive learning is easier and takes less time, it might be the best approach given the vast number of words that need to be learned. The results of Laufer and Shmueli's (1997) lead to similar findings. Results indicated that words glossed in L1 were better retained in long-term memory than those glossed in the L2. Laufer and Shmueli also suggest that mental elaboration might not occur when learners encounter L2 words in texts, however, bilingual lists can result in this kind of elaboration.

Taken together, the results of the three studies support the use of deliberate vocabulary study techniques that offer affordances for receptive and productive learning. The results showed that studying words receptively and productively are both effective, but receptive study produced superior results overall, is faster and easier, and it might not be not necessary for learners to do both. However, in a natural learning context, most learners first learn a word receptively through encountering it in a text, and then might choose to study these words deliberately, most likely through a combination of both receptive and productive learning.

Summary of the Studies Concerning Writing Versus Reading Tasks and Language of Presentation

Table 5 summarizes the main finding of the studies concerning both types of receptive and productive learning methods: (a) writing versus reading, and (b) learning from the L2 to L1 or from the L1 to the L2. As seen in Table 5, while both types of receptive and productive learning produce significant vocabulary gains in both types of knowledge, both forms of productive learning take more time and are viewed as being more difficult. In terms of reading activities (receptive) and writing activities (productive), again, both learning techniques proved to be successful in increasing vocabulary in general, and also lead to gains in both types of knowledge.

However, with the exception of Barcroft (2007b), the results of the studies indicate that productive writing tasks result in more gains in productive vocabulary knowledge than the receptive tasks and lead to similar levels of receptive knowledge. However, productive tasks take longer, so if time-on-task is a primary concern, then the efficacy of productive tasks over receptive ones must be considered carefully. In experiment 2 of this study, the premade card treatment only requires receptive study (reading), but the self-made card treatment requires both receptive and productive tasks (reading and writing), albeit the writing element consists of simply copying of sentences.

Table 5. *Results of Studies Concerned with Language of Presentation and Writing Versus Reading Tasks*

Productive (Writing) Versus Receptive (Reading) Tasks	
Productive tasks take more time (more difficult)	Barcroft (2004, 2006, 2007b); Behzadi & Haji Pour Nezhad (2014); Hulstijn & Trompetter (1998); Webb (2005) Webb, & Piasecki (2018)
Both types of tasks result in significant gains in both types of test scores	Behzadi & Haji Pour Nezhad (2014); Hulstijn & Trompetter (1998); Webb (2005), Webb, & Piasecki (2018)
Productive tasks result in better learning outcomes if time-on-task does not have to be equal	Hulstijn & Trompetter (1998); Webb (2005), Webb, & Piasecki (2018)
Productive tasks result in lower learning outcomes than receptive tasks if time-on-task is equal	Barcroft (2004, 2006, 2007b); Webb (2005)
Productive (L1-L2) Versus Receptive (L2-L1) Tasks	
Productive tasks take more time (more difficult)	Laufer and Shmueli (1997); Mondria and Wiersma (2004); Waring (1997)
Receptive learning might lead to better outcomes than productive learning	Laufer & Shmueli (1997); Waring (1997)
Both types of tasks result in significant gains in both types of test scores	Laufer and Shmueli (1997); Mondria and Wiersma (2004); Waring (1997)

Concerning language of presentation, while all the studies found that both modes of presentation resulted in significant gains, the results generally show that receptive learning is superior in terms of learning outcomes, especially as it takes less time. Given the nature of word cards, students have the choice to study either receptively, productively, or using a combination of both. While this issue is not a specific focus of this research, qualitative data provide insights into whether learners more often study receptively, productively, or engage in both.

Learner Strategies

Moir and Nation (2002) carried out a qualitative study to examine what beliefs and strategies learners draw on to guide their vocabulary learning, and to identify how effective these strategies were. The main focus of the study concerned issues such as selection of words, increasing depth of knowledge, and strategy usage. As part of their English language courses, learners were taught about a variety of principles and strategies that could be used to maximize their vocabulary learning, for example, selecting words that were useful to their goals, utilizing a variety of strategies to enrich knowledge of target items, revising previously studied words, and carrying out activities that will help promote the ability to use the items productively. The participants were ten adult learners from seven countries who were enrolled on an intensive ESL course in New Zealand. Participants ranged in age from 20 to 32 and had been studying English for between two and 14 years. The study utilized semi-structured interviews as the main data collection method, but recent test scores and vocabulary notebooks were also used in order to help triangulate the data. The main findings were that although taught a variety of effective vocabulary learning strategies, nine out of the ten participants largely did not utilize them on a regular basis. One major finding most pertinent to this study was that participants did not use effective word selection strategies. The participants often selected words that were of limited use or little personal interest, and selection was often quite random. A second important finding relevant to this study was that most learners expressed a lack of interest in increasing depth of word knowledge, only learning aspects of word knowledge such as collocations and word families in order to pass weekly tests. Another major finding was that very few learners opted to use a variety of strategies to learn target items, instead opting for repeated rote learning. Only one participant used strategies such as the keyword technique, actively trying to seek extra information about the words, or trying to use them creatively in spoken or written form. While the above findings are very important and informative, one of the most concerning findings was that none of the participants regularly revised words learned in the previous week; participants simply studied for the weekly tests and then never revisited the words again. The repeated peer-generated test examined in Experiment three of this study aims to overcome this issue.

The findings of this study are very interesting and relevant to the current study described in this study. Although the study had a very limited sample size, the participants' responses intimate that even after considerable learner training, many students will ignore what they have been told in terms of the best ways to maximize vocabulary learning. The participants in the study largely stuck to a very limited study routine of rote learning, did not repeatedly revise learned words, and did not make informed choices in terms of vocabulary selection. Therefore, language educators need to find ways of aiding learners to efficiently select words that are useful to their goals, or at least useful in a variety of contexts where the L2 is used. Also, the results intimate that teachers also need to try and find creative ways to encourage learners to revise words over the long term, not just on a short-term basis in order to meet short term goals such as weekly tests. In addition, activities that promote depth of knowledge increments and creative use will also be of benefit to learners. The design of the word cards used in experiments one and two, along with the repeated peer-generated test examined in Experiment three of this study aim to overcome such issues.

Mizumoto (2010) carried out a study to investigate the vocabulary learning strategies (VLS) used by Japanese EFL university learners. The sample consisted of 139 18-22-year-old female university EFL learners majoring in humanities in a private university in western Japan. Mizumoto used a vocabulary learning questionnaire composed of two sections; section one included 17 questions on beliefs about vocabulary learning, and section two composed of 91 items concerning vocabulary learning strategies. When reporting the results, only descriptive statistics were presented, but it was stated that the mean scores reported below were significant at the $p < .05$ level. The main results in terms of student beliefs about vocabulary learning were that the learners felt words should be acquired in context ($M = 4.29$, $SD = 0.65$) and put to use ($M = 4.09$, $SD = 0.66$) rather than memorized by rote. This led Mizumoto to state that the results "imply that Japanese EFL learners feel they understand the importance of learning vocabulary in context and actually using it" (p. 64). In terms of strategies used, guessing strategies and dictionary use strategies showed higher scores than other VLS, and two repetition strategies (oral and visual) were more frequently used than memory strategies. The results led Mizumoto to conclude that "Among cognitive and memory strategies, repetition strategies are used more frequently compared with imagery and association strategies" (p. 64). In addition, note-taking strategies, a category that Mizumoto believes could involve making word cards, scored quite highly ($M = 2.91$, $SD = 0.97$), and received higher scores than the strategy of using word lists ($M = 1.93$, $SD = 0.98$).

While there were many other results and conclusions, the ones most relevant to this study are the fact that learners seem to have fairly high levels of awareness about what kind of strategies are available to them, but generally rely on a select few, something that is in line with the Moir and Nation (2002) study. Of the various VLS available, participants seemed to prefer guessing meaning from context, using a dictionary to find the L1 meaning, and carrying out some kind of productive (written) task like taking notes or making word cards or notebooks. Learners also reported regular repeated study, be it through reading information about the words' multiple times or through verbally using the words, however, little is known about the actual techniques used. Simply learning from lists appeared not to be highly regarded or regularly carried out based on learners' self-reports. These findings are relevant to the longitudinal part of the study described in experiment 1, as well as to the peer testing system used in Experiment 3, as both methods involve learners identifying unknown words, checking the L1 meaning with a dictionary, then writing down this and other information, and then repeatedly and regularly studying the words on their cards either individually or with a partner.

Kojic-Sabo and Lightbown (1999) investigated the vocabulary learning strategies used by ESL and EFL students and examined the possible relationship between strategy use and achievement level. The sample consisted of two groups, the first of which was an ESL group containing 47 undergraduate students at a Canadian university, and the second group consisted of 43 EFL students enrolled in their final year at a university in Yugoslavia. The independent variables were self-reported scores regarding the amount of time learners usually spent on vocabulary learning, the extent to which they engaged in independent language study, the type of vocabulary learning activities they did on a regular basis, the frequency and elaborateness of their note-taking and reviewing efforts, and the frequency and elaborateness with which they used dictionaries. The dependent variable was participants' performance on two tests; a Yes/No academic vocabulary test and a cloze test assessing overall English proficiency. Data were analyzed using a series of one-way ANOVAs and cluster analysis.

The most important and relevant results and conclusions drawn for this study were that (a) students' strategy use was quite limited overall, (b) strategy choice was quite different amongst even a small group of learners, (c) students spent very little time studying vocabulary outside of class, (d) ESL students might have more opportunities to unconsciously encounter, use, and acquire new words, while EFL learners have to put in a concerted and significant effort to achieve the same results. Furthermore, the strategy that was reportedly used most consistently by both groups was note taking, presumably taking time to make notes about new words or phrases as they are met in context.

This paper is relevant to the experiments described in this study for a number of reasons. First, due to the fact that students appear to spend very little time studying vocabulary out of class and largely rely on few strategies, it is important to have them use effective activities for the short time that they do spend. In addition, word card making could be described as a form of note taking, the strategy that was reported as being most commonly used amongst the participants in the above study. Therefore, if the study on word cards described in experiment 1 can show their effectiveness in terms of vocabulary learning outcomes, it would offer further support for their inclusion in language course curriculums. Second, in addition to the fact that learners reportedly spend little time studying outside of class, as EFL learners have less opportunities to meet new words naturally or without significant effort outside of class, activities that offer chance for further study and self-evaluation of learning could be beneficial, and it is hypothesized that the peer test system described in experiment 2 offers this.

In a similar study, Barcroft (2009) investigated the strategies used during intentional vocabulary learning and assessed the relationship between strategy use and vocabulary learning. The original participant sample consisted of 120 first-year native English-speaking students of Spanish at a private university in the United States, however, due to various logistical and research design issues, only 83 participants data were analyzed in the study. Learners took a 24-item passive recall pretest before studying the target words. Words were then presented at the front of the class, with each word being presented separately for six seconds on slides in a computer presentation program with a picture (line drawing) of the word referent above each word. The participants then completed two different recall posttests, one identical to the pretest, and the other being a picture-to-L2 recall test. A lexical production scoring protocol was used to score the test results in order to give points for partial knowledge, the scale of which was 0, 0.25, 0.5, 0.75, and 1. In addition, participants answered questions about their strategy use, and these responses were coded to identify each participant's most frequently used strategy (MFS) and frequency of use of each MFS was calculated. A subset of posttest scores was analyzed using repeated measures analysis of variance (ANOVA) with MFS as a between-subjects independent variable, recall type as a within-subjects independent variable, and vocabulary scores as the dependent variable. Only MFSs reported by at least 5 participants was submitted to analysis. The MFS reported most often was L2-picture association, followed by L2–L1 association, L2–L1 translation, repetition and mnemonic technique.

The ANOVA conducted on these five MFSs only revealed significant main effects for MFS, $F(4, 63) = 3.31$, $p = .016$, $\eta^2 = 0.173$, and recall type, $F(1, 63) = 36.15$, $p < .001$, $\eta^2 = 0.365$. Overall means were significantly lower for picture-to-L2 recall, $M = 11.91$, $SE = 0.58$, than for L2–L1 recall, $M = 14.56$, $SE = 0.66$. A significant positive correlation was also observed between the number of strategies used and picture-to-L2 recall score ($r = .328$, $p = .001$) and n L2–L1 recall score ($r = .231$, $p = .026$). Overall, Barcroft stated that the results indicated significantly better scores for mnemonic technique and L2–picture association over L2–L1 translation and repetition, and that higher frequency of use of MFS leads to greater learning. The main conclusions drawn by Barcroft are that (a) while many different strategies were reported, only a few of these were regularly used, (b) of the five commonly used strategies, mnemonic technique was the least frequently used, but the most successful, and (c) greater use of MFS leads to better vocabulary learning results.

The results of this study are important as they shed more light on what learners actually do when trying to memorize new words. The finding that learners seem to know about many strategies, but generally stick to a few core ones is in line with the studies by Kojic-Sabo and Lightbown (1999); Moir and Nation (2002), and Mizumoto (2010). The results again indicate that learners are aware of the need for vocabulary learning strategies but might need more training or instruction about using a wider variety of them. In addition, learners could benefit by being given access to teacher designed and lead activities that incorporate various vocabulary learning strategies.

In another study, Barcroft (2003) investigated the effect of asking questions about target word meaning, something referred to as semantic elaboration, while trying to deliberately learn new vocabulary. The participants were 20 native English-speaking university students enrolled in two classes of second-semester Spanish at a private university in the United States. The study utilized a within-subjects design and included two learning conditions: (a) semantic elaboration via addressing questions about word meaning while trying to learn new words (+semantic), and (b) a free study condition referred to as the 'do your best only' (comparison). The study followed a pretest-posttest design and included three tests that aimed to measure the learning outcomes of asking meaning focused questions while studying vocabulary.

During the treatment, the participants attempted to learn 24 concrete Spanish nouns (12 in each condition) while seeing the L2 word and a representative picture. Each word and picture pair were shown twice for six seconds each. In the +semantic condition, the participants answered two questions during each presentation of the words. The questions that the participants thought about during the first learning stage were *When was the last time (if ever) that you used this object? and In what ways can this object be used?* Then, during the second trial, they were shown the following questions: *Where can you buy this object* and *What other objects do you associate with this object?* For the comparison condition, participants were only told to do their best to learn the new words.

After the treatment, three tests were administered to assess the effect of each learning condition on L2 free recall, L1 free recall, and L2 cued recall. Therefore, after the treatment phase, an immediate L2 (Spanish) free-recall test was administered that required the participants to recall the Spanish words that they had recently seen in the exposure phase. After this, an English free-recall test was administered. Finally, after carrying out an unrelated classroom-based vocabulary activity for ten minutes (distractor activity), a second L2 cued recall posttest was administered where participants attempted to write the L2 word after being shown its representative picture.

Data for the Spanish and English free-recall scores were subjected to an analysis of variance (ANOVA) with condition (+semantic, comparison) and language (Spanish, English) being treated as within subject independent variables, and the free-recall score was the dependent variable. In addition, Spanish cued- recall scores were submitted to a separate ANOVA with measure (posttest and delayed posttest) and condition (+semantic, comparison) being treated as within-subject independent variables, and the cued-recall score was the dependent variable.

The results for the free recall scores revealed a significant main effect for language due to greater free recall of English words than Spanish words, but the effect of learning condition was not statistically significant. However, means score on all posttests for cued recall of Spanish words were higher in the comparison condition than in the +semantic condition. The results for the second ANOVA revealed a significant main effect for condition and for time with mean scores again being higher in the comparison condition. Overall, the results showed that addressing questions about word meaning during L2 lexical learning did not lead to greater L2 or L1 free recall than the comparison condition and led to less cued recall. Unfortunately, no statistical data are presented in the paper, so traditional data reporting in terms of statistical significance or effect size is not possible.

Barcroft concluded that semantic elaboration using meaning-focused questions is inhibitory, and the additional semantic processing needed for such activities distracts from learning of new L2 vocabulary. Barcroft went on to state that the results of the study appear to be consistent with the predictions of TAP theory and the TOPRA model in their predictions that the effect of increased semantic processing depends on the nature of a task and testing, and that increased semantic processing can inhibit performance on a task that is not semantically oriented in nature. He stated that the results of the study draw into question the effectiveness of questions about word meaning as a technique for L2 vocabulary instruction and highlight the possible negative effects of eliciting large amounts of semantic processing during L2 vocabulary learning. Overall, it is concluded that the results of this study intimate that leaners might better learn the meaning of new words when the focus is simply on L1-L2 form-meaning connections. Therefore, it is suggested that it might be best, at least in the early stages of learning, to have learners simply focus on the meaning (translation) of new words instead of asking them to make higher level sematic elaborations.

This study is pertinent to this research as it offers support for the use of word cards as they allow learners to focus on the L1 meaning of new L2 words. In addition, the results intimate that word cards are an effective method as they do not involve unnecessary semantic elaboration; learners do not have to answer any meaning focused questions during the study period. Furthermore, in terms of the peer testing system analyzed in this study, as this system involves relatively simple form-meaning recall questions, it can be hypothesized that it is an effective method of enhancing knowledge of previously studied vocabulary.

In another study, Lavoie (2016) carried out an experiment to investigate the extent that training on vocabulary strategy use results in increased usage of these strategies, and how this might affect vocabulary learning. The participant sample consisted of 54 Indigenous Canadian third-grade students learning French as a second language and who spoke English as a first language. There was a total of 27 girls and 27 boys from seven classes in three schools. The study involved a treatment group who was trained in three vocabulary strategies (word family strategy, word network strategy, and word card usage) over a number of weeks, and a comparison group that received no training and learned words mainly by wrote or through textbook based activities. Learners took a pre and posttest past recall test consisting of 15 words. After taking the pretest, learners were given time to complete three treatment-based tasks with each of the new words before taking the posttest. The data were analyzed using repeated measures MANOVA with the independent variable being treatment condition, and the dependent variables being amount of strategy use and word knowledge test scores.

The results showed a significant difference between the two groups for the three strategies, $F(3, 51) = 6.89$, $p < .01$, $\eta^2 = 0.29$. In addition, a significant time effect for the application of the three strategies was found, $F(3, 51) = 21.52$, $p < .01$, $\eta^2 = 0.56$, and the main effects were qualified by a significant interaction, $F(3, 51) = 18.24$, $p < .01$, $\eta^2 =$

0.52, suggesting that the changes in the pre and posttest results for the use of the three strategies were different for the two groups. While there was no significant difference between the groups on the pretest, posttest results showed significant differences; $F(3, 51) = 16.10$, $p < .01$, $\eta^2 = 0.49$. While there are many interesting results, all of which are supported by well reported statistical data (pp. 27–29), the main results were that the group who received training used all three strategies to a statistically higher degree than the comparison groups, and this strategy use positively affected vocabulary knowledge gains. Lavoie concludes that the results demonstrate the success of training learners in various vocabulary learning strategies, and activities that give them additional in class chance to practice these strategies will result in better learning than simple rote learning.

The results of this experiment are of interest to Experiment 3 described in this study because, based on observation data to date, the peer test system gives learners the opportunity to practice good vocabulary strategies such as shuffling their word cards, generating new sentences, and thinking about other word forms. In addition, it could be that it offers learners with a stimulus to assess the efficacy of their vocabulary study practices. For example, if they score low on the test, they might try to make changes to their study routines, possibly employing better practices or using a greater amount of strategy use.

Overall, the studies reviewed above have provided various interesting and pertinent insights in terms of vocabulary learning strategies, and the results share many similarities.

First, there were many similarities amongst the results regarding the range of strategy use. For example, Moir and Nation’s (2002) study highlighted that even when taught a variety of effective vocabulary learning strategies, learners do not utilize them on a regular basis and rely on only one or two techniques. Similar results were also found in Mizumoto (2010) and Kojic-Sabo and Lightbown (1999); learners mainly rely on a limited range of deliberate vocabulary study techniques such as using dictionaries and making word cards or vocabulary notebooks. Barcroft (2009) also found that while learners had knowledge of many different strategies, only a few of these were regularly used. In addition, due to the results of their experiments, Barcoft (2009), Kojic-Sabo and Lightbown (1999), and Lavoie (2016) also concluded that wider use of vocabulary learning strategies leads to better vocabulary learning results. Lavoie (2016) also showed that learner training can results in increased strategy use.

Next, some of the studies found similar results regarding motivation. Moir and Nation (2002), Mizumoto (2010) and Kojic-Sabo and Lightbown (1999) all found that learners had relatively low levels of motivation for vocabulary study. In addition, Kojic-Sabo and Lightbown (1999) showed that learners admitted to spending very little time studying vocabulary outside of class.

In addition to the above findings, Moir and Nation (2002) also found that participants did not use effective word selection strategies, often randomly selecting low-frequency words from reading passages. In addition, most learners had little interest in increasing depth of word knowledge and only learned the information needed to pass weekly tests.

Based on findings of all the above studies, it seems that it is essential for language teachers to design tasks that require learners to focus on high-frequency vocabulary items and use various strategies when studying them. Also, to encourage learners to use various strategies, it is necessary to train learners in such strategies and give time in class to practice them (Lavoie, 2016). In addition, given the general lack of motivation (Moir & Nation, 2002; Mizumoto, 2010; Kojic-Sabo & Lightbown, 1999) and effects of regular testing (Moir & Nation, 2002), vocabulary tests could play a significant role in increasing both overall study time and strategy use. Testing could also be used to increase depth of knowledge by regularly administering vocabulary tests that include measures focusing on more than just the form-meaning connection.

Table 6 presents the major findings of the studies reviewed on vocabulary learner strategies.

Table 6. *Summary of Major Findings of Studies on Vocabulary Learning Strategies*

Learner Strategies	
Learners do not use a range of strategies	Barcroft (2009); Kojic-Sabo and Lightbown (1999); Moir and Nation (2002); Mizumoto (2010)
Learners rely mainly on deliberate learning techniques	Barcroft (2009); Kojic-Sabo and Lightbown (1999); Moir and Nation (2002); Mizumoto (2010)
Wider variety of strategy use results in superior learning outcomes	Barcoft (2009); Kojic-Sabo and Lightbown (1999); , Lavoie (2016)
Learners have relatively low levels of motivation for vocabulary study	Kojic-Sabo and Lightbown (1999); Moir and Nation (2002); Mizumoto (2010)
Testing results in additional study, and learners study only what is needed for the tests	Moir and Nation (2002)
Learners do not always focus on high-frequency words first	Moir and Nation (2002)

Vocabulary Self-Selection

The first stage in learning any new vocabulary item is noticing that a word or phrase is unknown (Nation, 2013, p. 103; Schmitt, 2008; Schmidt, 1990). Then learners will decide whether to take action to learn the word or not. For example, during reading activities, some words might not be deemed important enough to warrant deliberate study; learners might decide not to stop reading and simply skip the word or try to guess its meaning from context. Alternatively, they could stop reading and consult a dictionary to learn the word meaning before continuing to read the passage. Choosing to study unknown words encountered in a reading passage is probably one of the most common vocabulary self-selection strategies implemented by L2 learners. Another common and recommended method in L2 courses is letting learners choose target items from high-frequency word lists (Coxhead, 2014, pp. 3–4; Nation & Waring, 1997, pp. 6–19; Ruegg & Brown, 2014).

The Vocabulary Self-Collection Strategy

While the above paragraph discusses vocabulary self-selection strategies in general, some researchers have put forward more specific and planned reading-based self-selection strategies, the most prominent being the Vocabulary Self-Collection Strategy (Haggard, 1986). This strategy involves learners selecting words encountered from course textbooks or real-world printed materials such as posters or magazines. After noticing and selecting unknown words, learners then add them to their individual word list, and in some cases to an in-class *Word Wall* (Haggard, 1985). Learners then discuss and explain their selected words; explaining their meaning to classmates and stating where they found them and why they chose them. This student-centered activity has been reported to increase involvement, motivation, and learning outcomes (Abrams, & Walsh, 2014; Ruddell, & Shearer, 2002). However, limitations of this system include the fact that it has largely only been implemented and researched with children learning L1 words, and that it cannot be easily applied to non-reading courses, especially those in EFL contexts. In addition, as there is no control over which words students choose, there is no guarantee that the selected words are useful ones; they might not be high-frequency words (Reynolds & Shih, 2019).

L2 Vocabulary Self-Selection in Reading Classes

Autonomously choosing which words to learn is one of a range of L2 vocabulary learning strategies, and "the ability to make strategic decisions about learning at task-level is likely to be a determinant of effective learning" (Moir & Nation, 2002, p. 15). Vocabulary self-selection strategies can be most easily implemented in L2 reading classes, where learners can be encouraged to self-select and then study unknown words found in reading passages. See Reynolds and Shih (2019) for an example of how this can work in an EFL context. Written passages present a context for unknown words that helps learners make form-meaning connections (Rott & Williams, 2003), and self-selecting words from reading material is beneficial because it allows individual learners to focus on vocabulary that meets their own needs, and increases learner autonomy and intrinsic motivation (Moir & Nation, 2002).

However, language teachers need to be careful to provide texts that contain a suitable amount of high-frequency words, something which is not always easy to do. For example, Ruegg & Brown (2014) investigated the vocabulary content of a range of EFL textbooks and found that the average textbook did not contain enough vocabulary items for significant learning, and many of the words found in the texts were low-frequency words. In a similar study, Laufer (2014) examined some common EFL reading textbooks and concluded that due to the small vocabulary gains reported in the literature on incidental vocabulary learning studies, the number of words presented in textbooks is not enough to provide significant vocabulary learning opportunities. These findings are especially significant in language programs such as those found in Japanese universities, where textbooks are often the main source of vocabulary input for courses lasting between 15 and 30 weeks. For example, Laufer's (2014) study found that the average textbook provided around 200 target words, so if such a book was used for a one-semester university course in Japan, learners would be presented with less than 15 words per week, which is clearly not enough to significantly increase language ability (Nation, 2013, p. 39).

L2 Vocabulary Self-Selection in Non-Reading Classes

In non-reading classes such as speaking and writing classes, learners might not be provided with enough reading material to self-select a suitable amount of target words from texts. In such classes, the traditional method of providing learners with unknown target vocabulary items is often to rely on textbooks or teacher-selected vocabulary.

However, relying solely on these kinds of sources for vocabulary input has a number of disadvantages similar to those discussed in terms of using commercial reading textbooks. For example, such a system does not offer learners the chance to implement self-selection strategies and again, the number of target items presented in textbooks is often not enough to provide significant learning outcomes. Laufer (2014) examined the number of target words presented in various EFL integrated skills textbooks and found that only 39 new words can be learned from an average textbook. Similar findings were found by Ruegg & Brown (2014). To overcome these problems a commonly used and effective self-selection method in non-reading classes is encouraging learners to choose target items from frequency-based word lists that give learners the opportunity to select and focus on high-frequency words that they do not know (Coxhead, 2000, 2014, pp. 2–4; Moir & Nation, 2002).

There are various word lists that present words in order according to their frequency within a large language corpus. The most widely used of such lists are the General Service List (GSL), Academic Word List (AWL), New General Service List (NGSL), New Academic Word List (NAWL), British National Corpus (BNC), and the Corpus of Contemporary American English (COCA). While the GSL and NGSL deal with words which are used in daily life, the AWL and NAWL focus on words commonly found in academic sources, such as academic papers, books, and lectures. The GSL includes around 2,000 word-families, the AWL has 570, the BNC has 20,000, and COCA has 8,000. The NGSL contains around 2,800 modified lemmas, and NAWL around 1,000. While not important to this study, the distinction between a word family and a lemma is worth noting. A word-family includes derivatives, for example, *perfectly* and *perfection* are derivatives of *perfect* and they are one word-family. By contrast, a lemma includes only conjugated words, for example *go* conjugates *goes, went, gone, and going* and they are one lemma.

Regardless of the system used to build these word lists, they are valuable resources for learners wishing to self-select word to study. If learners self-select from these lists, and then deliberately study them, considerable learning can take place. Such a system involves learner autonomy and motivation, noticing, productive and receptive study, and negotiation of form and meaning, all of which are important factors in developing vocabulary knowledge (Nation, 2013, pp. 102–114).

In summary, as this review of the self-selection literature shows, there is ample theoretical and empirical support for adopting a vocabulary self-selection approach in L2 courses. Vocabulary self-selection is widely believed to be a motivating approach that offers superior learning outcomes than more teacher-focused approaches (Cobb, 1999; Masoudi, 2017; Nation, 2013, pp. 98–101; Ogawa 2012). Self-selecting target vocabulary is beneficial because it increases learner autonomy and allows individual learners to focus on vocabulary that meets their own needs or goals, something which increases intrinsic and extrinsic motivation (Moir & Nation, 2002; Ng & Ng, 2015; Ruddell, & Shearer, 2002). The most common self-selection strategy implemented by L2 learners is choosing to study unknown words encountered in a reading passage, however, how to make sure students are selecting appropriate words is a challenge with no clear solution (Laufer, 2014; Ruegg & Brown, 2014; Nation, 2013, pp. 37–40; Ogawa, 2012). For example, while the Vocabulary Self-Collection Strategy (Haggard, 1986) can be applied to the L2 classroom, it cannot be easily applied to non-reading courses, there is no control over which words students choose so there is no guarantee that learners will choose useful words. Therefore, language teachers need to be careful to provide texts that contain a suitable amount of high-frequency words, but this is not always easy to do given small amount and poor quality of vocabulary presented in many language learning textbooks (Ruegg & Brown, 2014; Laufer, 2014). In addition, in non-reading classes it might be difficult for EFL learners to encounter enough vocabulary items to make any real development. One solution to these problems is for teachers to ask learners to choose target items from high-frequency word lists (Coxhead, 2014, pp. 3–4; Nation, 2013, p. 38; Nation & Waring, 1997, pp. 6–19; Ruegg & Brown, 2014). This system ensures that learners are regularly studying useful vocabulary while still allowing autonomy of choice. Regardless of which method is used, allowing learners to self-select target vocabulary is seen to provide many benefits, especially if they are using good word selection strategies such as focusing on high-frequency words (Cobb, 1999; Nation, 2013, p. 38; Nation & Waring, 2013, pp. 6–19; Ruegg & Brown, 2014).

Benefits of Self-Selection Strategies

Regardless of which self-section approach is used, allowing learners to self-select target vocabulary is seen to provide many benefits. While much of the research on vocabulary self-selection has been carried out in L1 learning contexts (Abrams, & Walsh, 2014; Haggard, 1985, 1986; Harmon, Wood, Hedrick, & Gress, 2008; Ruddell, &

Shearer, 2002), there are numerous benefits of implementing vocabulary self-selection strategies in L2 courses, many of which are supported by empirical research.

First, one of the main reasons in support of the vocabulary self-section approach is that the vocabulary profiles of individual learners who have been streamed into language courses based on standardized test scores such as the TOEIC or TOEFL often differ greatly (Hirai, 2014; Iso, 2012; McLean, Hogg, & Kramer, 2014; Willis & Ohashi, 2012). The vocabulary self-selection strategy can overcome this challenge by allowing learners to focus on vocabulary suitable to their language level (Moir & Nation, 2002).

Furthermore, many mainstream language textbooks do not provide enough vocabulary items for learners to be able to make the kind of gains needed to increase second or foreign language mastery (Chow, 2007; Laufer, 2014; Ruegg & Brown, 2014). Therefore, self-selection from materials other than textbooks can increase the amount of vocabulary learners are presented with.

In addition, self-selecting target vocabulary is beneficial because it allows individual learners to focus on vocabulary that meets their own needs or goals, something which can increase intrinsic and extrinsic motivation (Moir & Nation, 2002; Ng & Ng, 2015; Reynolds & Shih, 2019; Ruddell, & Shearer, 2002). Also, allowing self-selection of vocabulary increases learner autonomy, another factor attributed to increasing motivation and learning outcomes (Dörnyei, 2001, pp. 101–109; Ryan & Deci, 2000; Moir & Nation, 2002; Ng & Ng, 2015). As noted by Reynolds and Shih (2019), if teachers select the target items instead of allowing the learners to do so, the learners might not feel ownership and might become demotivated, a viewpoint also expressed by other researchers (Adapa, 2015; Tseng & Schmitt, 2008).

In summary, the self-selection approach is better suited to mixed ability classes, and because students have more autonomy and freedom under the self-selection approach, it can be more motivating and increase engagement (Adapa, 2015; Schmitt, 2008; Tseng & Schmitt, 2008). Thus, it has been suggested that allowing students to self-select target vocabulary items is more beneficial than having a teacher or textbook prescribe what vocabulary all students should study (Harmon et al., 2008; Hirai, 2014; Laufer & Hulstijn, 2001). Therefore, many teachers allow learners to self-select the words they feel they need to study instead of assigning the same target vocabulary set to all students.

Motivation and Language Learning

Motivation is broadly defined as the stimulant for achieving a specific goal or target (Ryan & Deci, 2000). Motivation is widely accepted as being one of the most key factors of success or failure in L2 learning at all developmental stages (Dörnyei, 2001, p. 1; Hadfield & Dörnyei, 2013, p. 1; Harackiewicz & Sansone, 2000, p. 79). As Tseng and Schmitt (2008) stated "motivation has been widely embraced by both practitioners and researchers as a critical determinant of success in language learning, and this belief is strongly supported by a wide range of studies on L2 motivation in the past three decades" (p. 358).

Motivation provides learners with the primary momentum to initiate L2 learning (Ng & Ng, 2015, p. 1). Schmitt (2008) further states that without sufficient levels of motivation, regardless of ability or context, learners will be unable to accomplish long-term language acquisition goals. However, each learner has different beliefs and is motivated differently and to a different extent (Dörnyei, 2001, pp. 67–68) For example, some learners thoroughly enjoy the language learning process while others do not, some have clear external goals that can only be met by mastering an L2, while many see no clear purpose or benefit of learning the target language. In addition, even when dealing with initially motivated learners, it can take a lot of time and effort, both on the part of learners and teachers, to maintain sufficient levels of motivation (Ng & Ng, 2015; Schmitt, 2008). When discussing motivation, two distinct types of motivation have been identified; intrinsic and extrinsic motivation.

"Intrinsic motivation can be found within the individual and is related to the individual's identity and sense of well-being" (Ng & Ng, 2015, p. 1). In the field of SLA, intrinsic motivation is defined as the extent to which a learner works or strives to learn a target language because of a desire to do so, and because of the satisfaction the learning activities provides (Dörnyei, 2001, pp. 9–12). Put simply, learners are intrinsically motivated when learning is the primary or only goal. Intrinsically motivating tasks are enjoyable and challenging, and the reward is the enjoyment or feeling of competence or accomplishment that the task provides. While intrinsic motivation is without doubt the most important type of motivation (Ng & Ng, 2015; Ryan & Deci, 2000), extrinsic motivation also exerts positive effects on language learning and therefore has an important role to play (Deci, 1971, 1992; Dörnyei, 2001, pp. 9–12; Ng & Ng, 2015).

Extrinsic motivation comes from outside the individual; motivation is provided by an external reward or goal. Extrinsically motivating factors could be achieving good grades, receiving praise, being awarded a scholarship, or obtaining a job. These types of rewards are not directly associated with the learning itself; engaging in an activity and performing well is simply a necessity to earning those rewards. Learners who are extrinsically motivated will often continue to engage in an activity even if it is not interesting or rewarding in itself (Deci & Ryan, 1985, p. 35).

While intrinsic and extrinsic motivation are clearly very distinct constructs, it is widely asserted that they often interact and can affect each other (Dörnyei, 2001, p. 11; Harackiewicz & Sansone, 2000, pp. 79–84; Ng & Ng, 2015). Each learner can be seen to have a total motivation level, something which is made up of combination of extrinsic and intrinsic motivation (French, 1955; Walqui, 2000). Most researchers and scholars agree that intrinsic motivation is more closely related to success in learning a target language, but external rewards or goals can either increase or decrease intrinsic motivation (Workman & Williams, 1980). For example, learners who are motivated to pass a paper-based language proficiency test might find test practice activities more enjoyable than learners who aim to improve their speaking ability.

In summary, both intrinsic and extrinsic motivation play important roles in language learning. However, while intrinsic motivation is viewed as the most important and beneficial type of motivation, not all students enjoy the language learning process, thus teachers can use extrinsically motivating factors such as tests and grades to ensure learners have at least some level of motivation.

Motivation and Vocabulary Learning

As with any learning situation, motivation also plays a vital role in vocabular learning. As Folse (2011), Nation (2013, p. 102), and Schmitt (2008) have stated, long-term vocabulary learning requires considerable time and effort, therefore, high levels of sustained intrinsic and extrinsic motivation are necessary for success. As noted earlier, learners need thousands of words needed to function in a foreign language (Hu & Nation, 2000; Laufer, 1989; Nation, 2013, pp. 14–16; Webb, 2020), but acquiring these words can be a daunting task for many language learners so sustained motivation, preferably intrinsic, is vital for success (Schmitt et al., 2017). Without sustained motivation, learners will not devote enough time to vocabulary learning, and might even give up entirely.

However, maintaining motivation, especially intrinsic motivation, for long periods of time is not easily achieved, and all learners are likely to go through peaks and troughs in terms of their intrinsic motivation levels (Deci, 1992, pp. 43–70; Dörnyei, 2001, pp. 21–23; Schmitt, 2008; Ng & Ng, 2015). Therefore, it is the responsibility of researchers and teachers to identify activities and practices that increase intrinsic motivation for language learning (Hadfield & Dörnyei, 2013, p. 1), but if that is not possible, activities that offer extrinsic motivation can also prove useful (Harackiewicz & Sansone, 2000, pp. 79–104; Ng & Ng, 2015).

Extrinsic motivation, which also exerts positive effects on language learning (Deci, 1971, 1992; Dörnyei, 2001, pp. 9–12; Ng & Ng, 2015), can play an especially important role in deliberate vocabulary learning because even strong supporters of such an approach (Nakata, 2008; Nation, 2013; Reynolds & Shih, 2019) do not claim that it is intrinsically motivating. While incidental vocabulary learning activities such as extensive reading, watching movies, or listening to music are said to be intrinsically motivating (Bamford & Day, 1998; Edy, 2014), deliberate vocabulary study activities are rarely seen as intrinsically motivating (Al-Murtadha, 2014; Thompson, 2014; Wilkinson, 2015). Deliberate vocabulary learning can be viewed as a form of rote learning, something which is rarely intrinsically motivating for learners (Mantiri, 2015; Sakai & Kikuchi, 2009). Some studies have specifically shown that learners do not find deliberate vocabulary study enjoyable or interesting (Agawa, Black, & Herriman, 2011; Wilkinson, 2015, 2017). As a result, although making deliberate vocabulary learning enjoyable is the key goal, activities that increase extrinsic motivation for such activities are also beneficial to language learners.

In summary, motivating learners to devote the time and effort needed to deliberate vocabulary study is essential due to the growing body of research showing (a) the positive effects of motivation on vocabulary development (Tanaka, 2017; Workman & Williams; 1980, Zheng, 2012), and (b) the efficacy of deliberate vocabulary strategies (Cobb, 1999; Laufer & Rozovski-Roitblat, 2014; Nakata, 2008; Waring, 2004). Therefore, teachers must find ways of increasing intrinsic motivation for deliberate vocabulary learning methods such as word cards, for example, by allowing self-selection of target items and providing supplementary enjoyable deliberate learning activities. However, as it is difficult to make deliberate vocabulary learning activities intrinsically motivating, increasing extrinsic motivation could also play a vital role. To this end, given the known effects of testing on extrinsic motivation, along with other benefits discussed below, including regular testing into language courses is one viable option.

Test-Enhanced Learning

One common method of increasing extrinsic motivation in many learning environments, including foreign language courses, is testing students on the course content they have studied (Dörnyei, 2001, pp. 90, 134; Ng & Ng, 2015, Wilkinson, 2015). Tests are a popular method of increasing extrinsic motivation because it is believed that most learners enrolled on a course are motivated to get a good grade (Dörnyei, 2001, p. 90; Ng & Ng, 2015; Roediger & Karpicke, 2006a, 2006b). Repeated testing is especially likely to increase extrinsic motivation and study time because as Roediger and Karpicke (2006a) stated, if students know they will be tested regularly, instead of just once at the end of a course, they will study more and space their studying over the semester rather than cramming just before the final exam.

In addition, due to motivational and cognitive benefits, several researchers have shown that regular testing can effectively increase learning outcomes, especially long-term retention of target knowledge (Butler, 2010; Mcdaniel et al., 2007; Roediger & Karpicke, 2006a, 2006b; van den Broek et al., 2013). Results suggest that taking tests on previously studied material produces better retention over time than simply restudying the same material for an equivalent amount of time. A teaching approach that utilizes regular testing instead of simply restudying the same material again is often referred to as test-enhanced learning, and the positive learning outcomes are often termed the testing-effect (Butler, 2010; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a). Therefore, regular testing is a commonly adopted method of not only assessment, but of increasing learning outcomes.

In one of the only studies that examined this specifically from a second language vocabulary point of view, van den Broek et al. (2013) showed that including regular testing could improve learners' retention of previously studied vocabulary. The researchers showed that tests requiring memory retrieval strongly improve long-term vocabulary retention in comparison to continued studying. Once learners know the translation of a word, restudying the word by seeing the L1 and L2 forms together again is less effective than testing practice that only allows learners to see the word and then retrieve the translation from memory. Recall of the translations on a delayed posttest after one week was significantly better and faster for tested words than for restudied words. Positive results of testing over restudying were attributed to processes related to targeted semantic elaboration and selective strengthening of associations between retrieval cues and relevant responses. Similarly, Butler (2010) stated that repeated testing with different questions increases the probability of future retrieval by creating multiple retrieval routes in memory, and as a result, might also increase the probability of successful transfer.

The testing-effect is likely due to increased extrinsic motivation, repetition, and levels of engagement and processing because testing offers opportunities for self-evaluation, reflection, and revision (Butler, 2010; Deci, 1971; Roediger & Karpicke, 2006a, 2006b).

In terms of the connection between vocabulary development and testing, given the positive motivational and learning outcomes of test-enhanced learning, and because it provides additional opportunities for repetition, retrieval, and recall, regular testing can be hypothesized to be beneficial (Nakata, 2020, pp. 305–307). If learners know that they will be regularly tested on target vocabulary, they will likely be more extrinsically motivated to study these items regularly, will meet and recall these words more frequently during both the study and testing stages, and should therefore attain better vocabulary learning outcomes. Nation (2013) stated that tests can be an important part of the vocabulary component of a language learning course but stipulated that testing should not take up too much time (pp. 478–480). In addition, it is a commonsense notion that too much testing can reduce intrinsic motivation because many learners do not enjoy taking tests and can experience test anxiety. Therefore, it is preferable that a regular test system can be administered quickly and should be viewed positively by learners.

Self-Selection, Motivation, and Vocabulary Testing

While the vocabulary self-selection approach offers the potential to increase learner intrinsic motivation (Harmon et al., 2008; Masoudi, 2017, Moir & Nation, 2002; Reynolds & Shih, 2019) and regular vocabulary testing is likely to increase extrinsic motivation (Butler, 2010; Ng & Ng, 2015; Roediger & Karpicke, 2006a, 2006b), the two approaches are at odds with each other. This is because allowing vocabulary self-selection creates almost overwhelming challenges regarding testing students on their vocabulary development during a course. If students self-select vocabulary items, then each student has a different vocabulary set written on their deliberate study materials, especially given the differing vocabulary profiles of individual students within a class (Brown, 2012; Hirai, 2014; Iso, 2012; McLean et al., 2014). For this reason, standardized tests cannot be used, so it is difficult to provide learners with a method of assessing their vocabulary knowledge and growth. The inability to offer vocabulary tests for learners who self-select target items also leaves teachers with limited ways to assess students on their vocabulary development or increase extrinsic motivation. Due to this, it currently seems very difficult for teachers to adopt a vocabulary self-selection approach and regularly test learners on their vocabulary development. Therefore, even though both self-selection and testing offer distinct motivational benefits, amongst other benefits, many teachers might have to give up one in favor of the other.

Gaps in the Literature

The first major gap addressed in this study is the lack of research showing how much learning occurs from making word cards. As mentioned earlier, while word cards are widely used and many experts are ardent supporters of word cards, no research was found that investigated the effects of making word cards. Therefore, this research investigates the amount of learning that takes place simply through the action of creating word cards. As previously mentioned, there is currently no empirical data showing the amount of learning that takes place from the action of making word cards, or how sensitive this knowledge is to attrition if no further study is carried out. While studies such as those carried out by Behzadi and Haji Pour Nezhad (2014), Laufer and Rozovski-Roitblat (2014), Mondria and Wiersma (2004), and Webb (2005) examined some of the actions involved in making word cards such as writing information, retrieving information receptively or productively, and reading translations, no researchers have focused on the learning outcomes as a result of the specific process of making word cards. As a result, there is no empirical data showing how much vocabulary is learned from simply making word cards, or how robust this knowledge is over time.

The second gap in the word card literature that this research addresses is whether it is better for learners to make their own word cards or to simply study from pre-made cards. Nation and Webb (2011) stated that "making 50 cards can take half an hour or more" (p. 42), but according to a pilot study carried out by the researcher, if learners include sentences and other information on the cards or in notebooks, it takes an average of about one minute per card or entry. Based on the same pilot study, when given one minute per card, learners can study each card at least two or three times. Therefore, whether learners are better served by spending the considerable amount of time needed to make deliberate study materials such as word cards or simply studying from ready-made versions of the same materials has not been examined empirically. As a result, when time-on-task is controlled, it is not known if the amount of learning that takes place from making word cards justifies the amount of time it takes. For example, it could be that the affordances offered by the design or characteristics of word cards combined with multiple repetitions or meetings is most important, not the action of making them. If this is the case, it is better for learners to be provided with pre-made materials and to focus their attention on studying from them. However, if the action of making word cards increases the level of involvement, noticing, engagement, motivation, and ownership as currently thought, all of which produce superior learning outcomes (Laufer & Rozovski-Roitblat, 2014; Newton, 2013; Nation, 2013, pp. 102–114), then this research will offer the empirical support for this common practice. L2 vocabulary researchers have most often compared two or more study methods (Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014; Nakata, 2008), or compared differing levels of involvement (Mondria & Wiersma, 2004; Pichette et al., 2011; Webb, 2002; Yoshii, 2014), but no research was found that isolated the effects of making word cards. Others, such as Nakata (2008), Komachali and Khodareza (2012), and Kuo and Ho (2012), have specifically compared the efficacy of word cards with other study methods, but it remains unclear if the participants made the cards or not. Waring (1997) is one of the only researchers to examine word cards under two different conditions, but his focus was not on how the cards were made, but on the difference in learning outcomes as a result of learning from the L1 to L2 as compared to that of going from the L2 to L1. No researchers to date have used the same deliberate study method, in this case word cards, with the only independent variable being the way the materials were created; by the learner or by a third party. Overall, the studies mentioned above offer more support for learners making their own cards before studying from them as the process offers opportunity for both receptive learning (L2-L1) and productive learning (L1-L2), and the action of making cards is likely to increase involvement, but empirical research examining this issue is badly needed.

The third major gap addressed in this study concerns the issues of vocabulary self-selection and testing; how can teachers reliably test learners who are self-selecting their target vocabulary items? It has been argued that allowing learners to self-select the vocabulary items to be learned is superior to having the teacher provide the same set of items to a whole class (Harmon et al., 2008; Hirai, 2014; Mizumoto, 2013; Nation, 2013, pp. 37–39, 98–101). The self-selection system might be more effective for a number of reasons; (a) textbooks do not offer enough vocabulary items to constitute significant learning (Kaneko, 2013; Laufer, 2014; Ruegg & Brown, 2014), (b) the system is fairer as it allows for the fact that individual learners in any given class are likely to have different vocabulary profiles (Al-Murtadha, 2014; Brown, 2012; Iso, 2012), and (c) such a learner-focused system logically increases noticing, motivation, and self-regulation (Hirai, 2014; Schmitt, 2008). However, the self-selection system opens up challenges in terms of vocabulary testing that have currently not been addressed in the literature. Therefore, this study addresses the gap of how teachers can regularly test learners who are self-selecting their target vocabulary items. If teachers want their students to self-select vocabulary to be studied, then it becomes almost impossible for teachers to test individual students regularly on their vocabulary knowledge, even though they might want to, or are required to do so. In many courses, regular testing is an expected or preferred part of the curriculum, as it provides a means of checking learners' progress and assigning grades (Butler, 2010). However, if students self-select target vocabulary, it is almost impossible for teachers to able to test all students regularly on their vocabulary knowledge because through the process of self-selecting, learners each have different sets of target vocabulary items depending on their background and initial vocabulary profiles. Therefore, they cannot be regularly tested using traditional tests such as the VLT or VST, as these instruments are too blunt, and do not have enough versions to be used regularly throughout a course. In addition, it is logistically impossible for teachers to create individual tests for each learner in a class. However, no researchers have introduced and validated a clear, simple, and time effective means of testing individual learners' vocabulary development repeatedly during a course. Therefore, empirical evidence supporting a system that can overcome this challenge, such as the one used in this research, is needed. If such a system can be introduced, explained, and supported through research, it can prove a valuable tool in terms of not only classroom vocabulary Therefore, research is needed that introduces and validates a test system that allows learners to be regularly tested on their vocabulary development, even when they are self-selecting their target items.

Furthermore, this research examines the gap in the literature regarding the motivational outcomes of regular vocabulary testing in the language classroom. As noted by Schmitt (2008), Thompson (2014), and Tseng and Schmitt (2008), motivation is a key factor in long-term vocabulary development, because without sustained, long-term motivation, it is unlikely that learners will maintain the momentum and effort needed to master the thousands of words needed for successful second language use. While regular testing is believed to have positive motivational effects (Butler, 2010; Ellis, 2002;

Fulcher & Davidson, 2007, pp. 222–224), no research was found that investigated this in relation to motivation for deliberate vocabulary learning. Specifically, no research has been found that examined how the inclusion of peer-generated vocabulary tests affect motivation for deliberate vocabulary learning.

The various experiments involved in this study also address gaps caused by the research design of many previous vocabulary studies with regards the measurement of (a) vocabulary knowledge and (b) motivation. To begin with, experiments 1 and 2 aim to fill the need for word card studies that use more than one measure of vocabulary learning outcomes, and which take into account partial knowledge gains. As noted by researchers such as Laufer and Rozovski-Roitblat (2014), Nation and Webb (2011, p.102), and Webb (2009), many vocabulary researchers have assessed vocabulary knowledge with one dependent measure, for example active or passive recall, thus such studies have not provided accurate accounts of learning outcomes. Also, as noted by Barcroft (2007a), many researchers have only scored test answers as either correct or incorrect, thus have not measured partial vocabulary knowledge. For example, while learners might not be able to recall the word correctly, they might have learned something about the word that constitutes not only learning in itself, but also shows that they are on an incremental path to fully knowing the word (Barcroft, 2007a; Webb, 2009). Therefore, to address this gap, the vocabulary studies discussed in this study use at least two measures of word knowledge and utilize a scoring system similar to that gives credit for partial knowledge. Next, Experiment 3 aims to address the gap caused by the overreliance on self-report data to measure motivation. There is a serious lack of studies that have used concrete and clearly operationalized measures of motivation and due to the current overreliance on self-report data, few examples of more accurate methods of measuring motivation for deliberate vocabulary learning are provided in the vocabulary literature.

In addition, this study addresses the current lack of vocabulary studies that combine carefully controlled cross-sectional studies with more ecologically valid longitudinal designs. While there are many cross-sectional vocabulary studies, and an increasing number of longitudinal ones, very few have combined both types of design. In general, many researchers use either carefully controlled cross-sectional (experimental) designs or less controlled longitudinal ones, but rarely both (Creswell, 2014, p. 12). As a result, we often have no way of knowing if the results found from cross-sectional treatments will hold true in real-life contexts over a longer period of time (Mackey & Gass, 2005, pp. 111, 161, 303). On the other hand, while longitudinal data better represent what happens in actual natural learning contexts, it is very difficult to control or assess the impact of variables not under investigation (Creswell, 2014, p. 12; Mackey & Gass, 2005 pp. 161, 303), so we cannot be sure if the learning outcomes were affected by unknown variables or variables unique to the particular research context. Therefore, there is a need for research designs that combine the two approaches and thus offer the benefits of both.

Finally, this study addresses the gap in the literature concerning a lack of mixed methods approaches to vocabulary research. Through adopting various qualitative research methods, this study will provide data about student's perceptions and actions related to deliberate vocabulary study and vocabulary testing, thus providing valuable and needed insights (Webb, 2020, p. 7).

The final problem is that few researchers investigating deliberate vocabulary study methods such as word cards have employed mixed methods approaches. While many vocabulary studies have reported to use a mixed-methods approach, they have largely involved very little qualitative data analysis, and few researchers have collected and triangulated data from multiple sources such as interviews, observations, questionnaires, and artifact checking (Creswell, 2014, pp. 200–209; Creswell & Clark, 2017, pp. 54–55; Hatch, 2002, pp. 93–133; Mackey & Gass, 2005, pp. 201–202). This is a serious problem as the insights that can be garnered from such research can highlight various factors that might have affected the learning outcomes (Creswell & Clark, 2017, pp. 54–55; Webb, 2020, p. 7).

Purposes of the Study

The first purpose of this study is to investigate how much learning takes place from simply making word cards. This research is the first known study to show how much learning takes place from making word cards, and to demonstrate how sensitive this knowledge is to attrition over time.

The second purpose of this research is to compare the efficacy of making and studying from word cards with simply studying from pre-made versions. Whether or not the time needed to make word cards is of more benefit than spending the same amount of time simply studying ready-made versions has so far not been addressed in the literature. Therefore, I compare both cross-sectional and longitudinal learning outcomes from making and studying word cards with that of studying from pre-made equivalents over the same time period. In addition, qualitative data aim to provide insights into students' perceptions of making and studying from word cards.

The third aim of this study is to share the results of reliability analysis of the peer-to-peer test system so teachers and researchers can decide if it is a suitable method of assessing learners' individual vocabulary development. If the student-generated test system used in this study can be shown to be a reliable vocabulary assessment tool, it could be very helpful for language teachers, especially those who support the vocabulary self-selection strategy.

Furthermore, this study aims to examine the effects of different test conditions on motivation for deliberate vocabulary study. This study aims to compare the motivational effects of (a) no testing, (b) regular informal peer-to-peer verbal testing with no course grade weighting, (c) the same peer testing system but with course grade points assigned to the test scores, and (d) more traditional written tests. Qualitative data provide insights into students' perceptions of each treatment and highlights their effects on motivation and time-on-task.

In addition, I aim to provide not only carefully controlled cross-sectional data, but also to compare such results with longitudinal data generated from more authentic treatment conditions. Finally, by supporting the quantitative data with a variety of qualitative data, I aim to offer detailed insights into the various individual and group factors that might have affected the outcomes. This aims to provide more detailed information into not only what happened but also how and why.

Overall, this research aims to answer a number of specific and important questions that have yet to be addressed in the second language vocabulary acquisition literature.

Research Questions

The following research questions for the three main studies involved in this study are provided below. The first set of research questions aims to assess the efficacy of the traditional word card creation method, while the answers to the second set will enable the comparison of three different word card creation methods: self-made no-choice, pre-made, and self-selected self-made. The research questions used in experiment 3 will provide data regarding the motivational outcomes of various vocabulary test conditions and assess the reliability of students as examiners of peer-generated vocabulary tests.

Experiment 1

1. How much initial learning occurs from making word cards?
2. How sensitive is this knowledge to attrition if no further study is carried out?
3. What are students' perceptions of making the word cards?

Experiment 2

1. How much initial vocabulary learning occurs from the two controlled word card treatments: premade card treatment and student-made no-choice treatment?
2. Are either of the two controlled word card methods significantly more effective for developing short-term vocabulary knowledge?

3. Which of the three word card methods is superior in terms of vocabulary retention after repeatedly studying target words over a number of weeks?
4. To what extent does the level of task-induced involvement load predict the learning outcomes discussed in questions 2 and 3?
5. How are the three experimental word card methods viewed by the participants?

Experiment 3

1. To what extent does the test system affect extrinsic or intrinsic motivation?
2. How is the peer-to-peer test perceived by learners and how does this affect the answer to research question 1?
3. How reliable are students as examiners?

CHAPTER 3

EXPERIMENT 1: LEARNING OUTCOMES OF MAKING WORD CARDS

As mentioned in the Literature Review chapter, word cards are a widely recommended method of intentional or language-focused learning of vocabulary (Elgort, 2011; Nakata, 2020, pp. 304–307; Nation, 2013, p. 437; Schmitt, 1997, pp. 199–227), but no research has been found that explores how much learning takes place from making word cards. Nation and Webb (2011) stated, "We know from experience that after making the cards, learners can recall from 30 to 70 percent of the meanings on the first run through the cards" (p. 42), but experimental data supporting this statement have not been found. While there is a growing body of research examining the learning outcomes associated with word cards (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014), none of the researchers measured the learning outcomes of making the cards without rehearsal taking place.

In addition, no studies have been found that use qualitative data to explore what students do when making word cards or how they feel about the process. As a result, it is unclear what practical, logistical, or affective issues are involved with word card making, nor how they affect learning outcomes.

Semantic Processes Involved in Experiment 1 Word Card Making Treatment

As discussed in the Methods, Discussion, and Limitations sections of this chapter, the word card making treatment used in this study does not represent a totally natural word card making process. The need to reduce confounding variables led to a word card making treatment with slightly reduced levels of processing, however, the treatment does involve many of the same physical and semantic processes.

For example, one perceived strength of making word cards, including the word card making treatment used in Experiment 1, is that the process involves both receptive (reading) and productive (writing) vocabulary learning processes (Mondria & Mondria-de Vries, 1994; Nation, 2013, pp. 450–451), both of which have been shown to be beneficial to vocabulary learning (Behzadi & Haji Pour Nezhad, 2014; Kuo & Ho, 2012; Mondria & Wiersma, 2004; Waring, 1997). When making the cards, learners receptively focus on form and meaning when reading L2 target items and their L1 equivalents, and then focus on form productively when writing L2 items onto their cards (Mondria & Mondria-de Vries, 1994; Nation, 2013, pp. 447–459). However, copying the words on the cards cannot be considered productive use as it does not involve recall or creative use of target items in sentences or longer texts (Barcroft, 2007b). While student-generated vocabulary notebooks also offer the chance for limited productive learning like word cards, premade vocabulary materials such premade digital flashcards do not offer the same affordance. In addition, due to the two-sided nature of word cards, the L1 and L2 forms are not written side-by-side like in notebooks, so it is possible for learners to have limited opportunities for passive or active recall when writing the cards. For example, when writing the L2 form of the word on the card, learners can recall the L1 meaning again during the process (before looking at the other side of the card, dictionary, or word list again).

In addition, the processes involved in the word card making treatment used in Experiment 1 involves similar levels of processing as a more natural word card making condition. The basic premise of the Levels of Processing model (Craik & Lockhart, 1972) is that learning activities induce varying stages or levels of semantic or cognitive analysis. While there are numerous types of processes involved, Craik and Lockhart grouped them into two levels. Level 1 involves shallow processing, which occurs when only the physical qualities of something are encoded, for example, how the letters of a word look (structural processing) or how letters or words sound (phonemic processing). Level 1 also often involves repetition in order to commit new knowledge to short-term memory. Level 2, which is called deep processing, leads to better recall and transference of knowledge to long-term memory as it involves more meaningful analysis of information such as semantic processing, elaboration, and making associations, for example, giving words a meaning or linking them with previous knowledge.

In Experiment 1, learners are involved in a level 1 process when they first receptively encounter the L2 target item in the learner dictionary. Then, they are engaged in Level 2 processing, albeit to a limited extent, when they make the form-meaning connection after looking at the L1 translation given in the dictionary. In addition, it is highly likely that learners will relate the target items to similar words with similar meanings, most likely high-frequency words previously learned. Learners also carry out further semantic processing when they look at the way the word is used in the example sentence. According to Craik and Lockhart (1972), seeing words in a sentence is beneficial because (a) the sentences help learners to develop an understanding how a word can be used, and (b) sentences are particularly beneficial because these kinds of stimuli are highly familiar to learners and as such are processed at a deeper level quickly and are better retained (pp. 676–678). Copying the words onto word cards provides learners with further shallow and slightly deeper levels of processing because this process introduces new cognitive and physical processes, and requires more attention, repetition, and time-on-task, all of which are believed to create stronger memory traces leading to better short and long-term memory.

In summary, a word card making treatment was designed to most closely represent the process learners would normally engage in, but the need for experimental controls led to a somewhat unnatural word card condition.

Research Questions

Based on the gaps in the literature discussed above, the following research questions aim to assess the learning outcomes attributable to making word cards. By conducting cross-sectional research, much-needed data about the amount of short-term vocabulary learning resulting from making word cards are produced. In addition, the delayed posttest data give a strong indication of how sensitive any initial learning is to attrition if words are not studied again in a relatively short time frame. Finally, the last research question is designed to shed light on affective and individual learner issues that could have influenced the learning outcomes and provide useful insights to language teachers interested in having students use word cards for deliberate vocabulary learning tasks. The research questions that guide Experiment 1 are:

1. How much initial learning occurs from making word cards?
2. How sensitive is this knowledge to attrition if no further study is carried out?
3. What are students' perceptions of making the word cards?

Methods

Research Context

This experiment was carried out over one semester in two private universities—one women's university and one co-educational university—located in central Tokyo. The first university from which the sample is taken is a private Japanese women's university in Tokyo that was founded in 1918. The university consists of 12 departments but mainly focuses on liberal arts education. According to the university's homepage, the student body consisted of 3,917 full-time undergraduate students and 87 graduate students as of April 2016. The 2018 *hensachi* rating of the university, a rating system based on entrance exam difficulty and university selectivity, was 55 (very selective), which means that considerably less than 50% of applicants are accepted. The university places importance on English education and as such, the students are generally required to take more credit bearing English courses than the minimum required by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), especially students enrolled in the language Sciences department. In addition, approximately 5% of the student body is enrolled on an intensive English program called the Career English Course. However, the majority of students are enrolled in English courses offered through their respective department's general curriculum, and common English language courses include first-year oral communication, first- and second-year reading, and first- and second-year writing. Based on pre-course questionnaire data, and after talking with long-serving staff members, some students are highly motivated with regard to learning English, but many are not particularly interested in gaining high-level English skills.

The second university from which participants were recruited is a private co-educational university in Tokyo that is considered one of the top private universities in Japan. According to the university's official website, the typically total enrolment is over 13,000 students enrolled in one of 11 research institutes, each offering a number of academic programs. The university prides itself on its international focus and English education programs, which the university website states are shown in the number of language courses, English-medium programs, and study abroad opportunities offered to students. According to the website, the university has offered English-medium programs for more than 60 years, has been recently chosen for a series of subsidy programs being implemented by MEXT to internationalize universities in Japan, and is renowned for encouraging Japanese students to study abroad. Therefore, it is believed that the participants from this university had relatively high motivation for studying English.

Participants

The participants were first- and second-year students recruited from the two private universities described above. Three groups of learners were taken from three intact classes at the woman's university and one group from a class at the other private university. In total, there were 82 participants at the start of the study, but 10 were removed because they were absent from one or more stages or they did not fully complete the treatment. This deletion resulted in a final sample of 72 participants (62 female and 10 male students) aged 18 to 21 years old. Three groups of learners from the first university made up approximately 75% of the participant sample, with two of the groups being first-year students and one group being second-year students. These participants were all Japanese female learners aged between 18 and 21. Approximately 70% of the participants were enrolled in the department of language sciences, and the remaining 30% came from a variety of other departments. English language proficiency varied considerable within each class, and none of the classes were streamed by language level. TOEIC scores ranged from approximately 250 to 600, and all participants had received at least six years of English as a foreign language instruction in junior and senior high school before coming to university. I was responsible for teaching the students in all three courses; however, all participants were also enrolled in other English courses taught by different teachers. All students were enrolled in courses working towards improving their general English proficiency in the four skills of reading, writing, speaking, and listening, and some wished to improve their scores on the TOEFL iBT, IELTS, or TOEIC tests. The sample represents a convenience sample because the participants were chosen predominantly due to ease of access; I was their English teacher during the duration of the study.

The fourth group of participants was taken from one intact first-year class at the second university described above. The sample consisted of 22 students from a variety of majors but who were all studying the sciences such as mathematics, technology, and engineering. These students were enrolled on a one-semester English for academic purposes (EAP) course that met twice a week for 90 minutes each session.

The four intact classes used in this study come under the following course names in the universities from which the sample was taken: Writing Skills 1A, Writing Skills 2A, Basic Communicative English (BCE), and EAP 1. A more detailed description of each of the courses from which the sample was taken is provided below.

***Writing skills 1A (University 1)*.** The Writing Skills 1A course was a compulsory English course offered to first-year students from the Department of Language Sciences in the spring semester. In the fall semester, all students who receive a passing grade for Writing Skills 1A continue on to the Writing Skills 1B course, and they then move on to the Writing Skills 2A and 2B courses in their second year. There were 109 students enrolled in this course, who were then divided into five classes. Only one class was streamed by English ability, with the top scoring 20 students on the department's English section of the entrance exam being placed into the same class. All other students were randomly assigned to one of the four remaining Writing Skills 1A classes, regardless of level. The course, which met once a week for a total of 90 minutes, was focused predominantly on academic writing skills, but it also included elements of vocabulary development, reading, and oral communication. The main writing focus was on academic essay writing, but the course was also designed to develop writing fluency by having students carry out fluency writing through a number of quick writing activities over the course of the semester. The original sample consisted of 19 female Japanese students taken from one intact Writing 1A course, all of whom were 18 years old. However, two students were removed due to being absent from one or more of the experimental stages and one removed for not completing the treatment correctly, three students were removed, leaving 16 participants from this group taking part in the study. The course from which the sample was taken was one of the non-streamed classes, with TOEIC scores ranging from 280 to 410. After teaching this course and talking with the students informally, the students can be described as relatively low level, with mixed levels of motivation for studying English. Most students have no intention to study abroad for an extended period of time, and they have no clear goal or reason for studying English. Based on pre-course vocabulary tests using the VLT, the participants likely had less than 70% knowledge of the first 2,000 high-frequency words, those contained on the General Service List (GSL), less than 40% of the third 1,000, and less than 20% knowledge of the words in the Academic Word List (AWL).

Writing skills 2A (University 1). The Writing Skills 2A course was a compulsory English course offered in the Spring semester for second-year Language Science students who had passed Writing Skills 1A and 1B. In the fall semester, all students who received a passing grade continue on to the Writing 2B course. There were 84 students enrolled in this course, who were then divided into four classes. Again, only the highest-level class was streamed by English ability, while all other students were randomly assigned to one of the four remaining Writing Skills 2A classes, regardless of level. The course, which met once a week for a total of 90 minutes, was focused predominantly on developing academic writing skills, but it also included elements of written fluency development, vocabulary development, reading skills, and oral communication. The main writing focus was on writing academic essays and developing basic secondary research and citation skills. The Writing 2A course sample originally included 23 female Japanese students aged between 19 and 20 (M = 19.3 years); however, two students were removed due to missing one or more of the stages of the experiment; thus, there are 21 participants from this intact class. Based on observations and informal interview data, the students have mixed levels of motivation for studying English, with most having no intention to study abroad and no clear goal or reason for studying English. In terms of language proficiency, their TOEIC scores ranged from 410 to 605 (M = 515). Students at this level are able to write short academic English essays similar to those required on standardized tests such as IELTS or TOEFL iBT; however, they have difficulty with organization, content, grammar, and vocabulary. Pre-course VLT scores indicated that the participants had less than 80% knowledge of the first 2,000 high-frequency words, less than 50% of the third 1,000, and less than 40% knowledge of the academic vocabulary.

Basic communicative English (University 1). The Basic Communicative English (BCE) elective English course is described by the university as a remedial English course offered to first-year students from any department in the spring or fall semesters. Students who are deemed as needing additional help in developing their English proficiency are offered the opportunity to take this course for one semester. Of the students identified as needing this course, based on the English section of the university entrance exam, approximately 50% of the lowest proficiency students can take the course in the Spring semester, while the remaining students can take it in the Fall semester. The course meets twice a week for a total of 180 minutes and is focused predominantly on developing conversation and discussion skills. As a means of achieving these goals, the course includes elements of vocabulary development, grammar teaching, listening practice, and speaking practice. The BCE course originally contained 18 female Japanese students aged 18, but one participant was removed due to missing the treatment stage and one for not completing the treatment fully, leaving 16 participants from this course. Based on observations and informal interview data, these students can be described as having fairly high levels of motivation for studying English, with most students stating that improving their English ability is one of their main goals during their university education. However, most of the students reported having no intention to study abroad for an extended period of time, and they had no clear goal or reason for studying English. In terms of language proficiency, their TOEIC scores ranged from 230 to 390 (M = 330). Based on pre-course vocabulary tests using the VLT, the participants are likely to have less than 70% knowledge of the first 2,000 high-frequency words, those contained on the General Service List (GSL), less than 30% of the third 1,000, and less than 20% knowledge of the words contained in the Academic Word List.

EAP 1 (University 2). The EAP 1 course is a one-semester compulsory English course offered to first-year students from any department in the private co-ed university in the spring semester. There were 80 students enrolled in this course, who were then divided into four classes. The course, which met twice a week for a total of 180 minutes, was focused predominantly on writing, reading, summarizing, and discussion skills, but it also included elements of vocabulary development. There were 22 students enrolled in this course, all of whom were aged either 18 or 19 (M = 18.20). However, due to missing one or more stages of the experiment or not completing the treatment, three students were removed from the study, leaving 19 students; 10 males and 9 females. Based on observations and informal interview data, the students had fairly high levels of motivation for studying English, with approximately 50% of the students stating that improving their English ability was one of their main goals during their university education. In addition, most of the students reported having an intention to study abroad for an extended period of time in the near future. However, as this was a compulsory class, the students took the course to reach the required number of English credits. Pre-course VLT scores indicated that the participants had knowledge of approximately 70% of the first 2,000 high-frequency words, but less than 50% of the third 1,000 and the words contained in the Academic Word List.

In summary, the original sample included 82 participants, but 10 were removed due to practical reasons. Seven students were removed as they were absent from one or more of the experimental stages, and three were removed for not fully completing the treatment. These three students were observed falling asleep; one during the treatment and two during one of the test stages. Thus, for analysis purposes, Experiment 1 had a total of 72 participants (62 female and 10 male students) aged between 18 and 21 who were enrolled in university EFL courses in one of two universities. Fifty-three students

(74%) came from three intact classes from the women's university, and the other 19 (26%) came from the private co-ed university. For this reason, 86% of the participants are female students.

Research Design

Experiment 1 is a mixed-methods study that utilizes a within-subjects, pretest-posttest design. The experiment utilized a cross-sectional design because the treatment took place only once over a short period of time, but the experiment does include a delayed posttest. The main data were quantitative, but qualitative data were also obtained from observations during the treatment stage and from eight post-treatment interviews.

The quantitative data were gathered from the three dependent measures of vocabulary knowledge; passive-recall, active-recall, and a cloze test. The quantitative data were used to measure vocabulary gains attributable to the treatment while the qualitative data obtained from observations and a small number of interviews offered further insights into what actions or processes might have influenced the results. However, as the qualitative data were not the main focus, and only eight interviews were carried out, Experiment 1 is a quantitative study rather than a mixed-methods design (Creswell & Clark, 2017, pp. 51–54). The research design is represented in Figure 1.

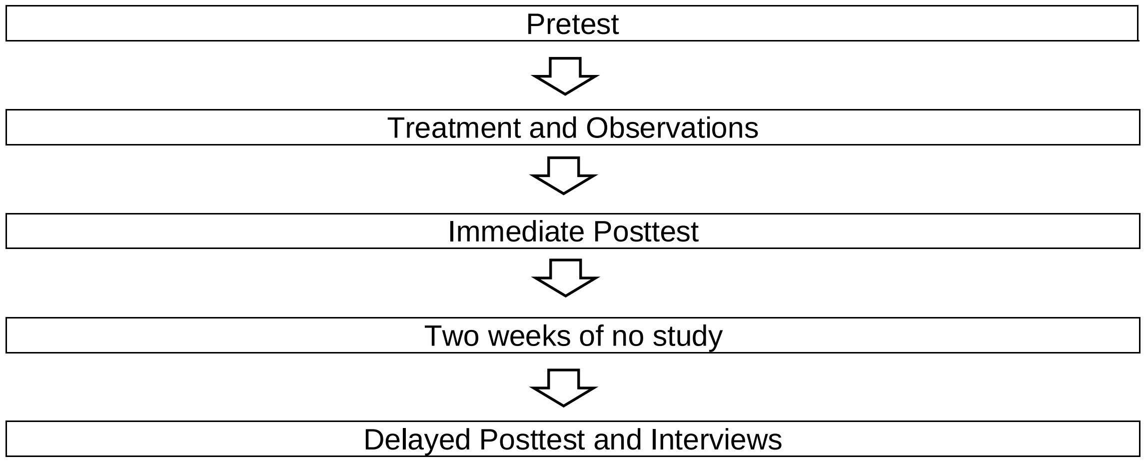

Figure 1. Research design for experiment 1.

As mentioned above, the participant sample included students from four intact classes from two universities. The fact that intact groups were used is usually a limitation as there can be no guarantee that the groups are equal on any number of variables, or that they are exposed to similar confounding variables during the course of the study. As Farrokhi and Mahmoudi-Hamidabad (2012) pointed out, when using intact groups, there is a high probability that the selected groups differ in many ways that do not allow the groups to be considered comparable. Issues such as group dynamics, time of the class, and other influences can affect the comparability of the groups. However, this study used a within-subjects design in that all students carried out the same treatment with the same words and tests; as a result, most of the problems associated with intact groups were eliminated. As Felix (2008) and Fendler (2016) state, when using a within-subjects design, each participant experiences the same treatment condition and acts as his or her own control, leading to increased reliability through the exclusion of some confounding variables that often negatively affect between-subjects designs with small intact groups.

Furthermore, a within-subjects design and the use of comparison items instead of a control group were used to increase the *ethical validity* of the study. While most criticisms regarding the use of control groups comes from the medical and psychological research literature (Conner, 1980; O'Leary & Borkovec, 1978), they can also be applied to educational research (Fendler, 2016). For example, using control groups that either do nothing or carry out a learning activity that is viewed as less effective than the educational treatment of focus denies those participants possible learning opportunities (Fendler, 2016).

Overall, the mixed-method approach aims to provide holistic and triangulated answers to the research questions, and the within-subjects design and use of comparison items increases the experimental sample size and ensures all participants have the maximum opportunity for learning.

Materials: Vocabulary Items

Selecting possible target items. The target items were initially chosen based on the results of the Vocabulary Levels Test (VLT) (Nation, 1983). First, as the VLT scores intimated that most participants did not have mastery of words at or beyond the first 2,000-word frequency level, it was decided that the participants would study words from the second and third 1,000 frequency levels of the NGSL as part of the course requirements. This study took place outside of this experiment.

The VLT was chosen to assess the participants' vocabulary knowledge due to its availability, ease and speed of administration, and the support for the VLT discussed by Kremmel and Schmitt (2017) and Schmitt, Schmitt, and Clapham (2001). The test is freely available, it can be administered in a relatively short time, and it has been shown to provide robust estimates of learners' vocabulary knowledge of words at various frequency levels. In addition, as part of normal course requirements, participants self-selected target words predominantly from the NGSL, which shares many words with the GSL, so the VLT scores allowed the students to focus their attention on a suitable word frequency level. The version of the VLT used in this study assessed the participants' vocabulary knowledge at two frequency levels, the second and third 1,000 words of English and the AWL. The reason that the 5,000 and 10,000 levels were not used was initially due to research estimating Japanese university students' vocabulary size. For example, Willis and Ohashi (2012), Nonaka (2004), and McLean et al., (2014) found that Japanese university students' vocabulary sizes are between 2,300 and 3,500 words on average. While some students had larger vocabulary sizes, many of the words known were not high-frequency words. In addition, pilot testing was carried out using the 5,000 and 10,000 VLT frequency levels with a group of learners at a higher English proficiency level than those participating in this study. The results showed that these learners knew less than 20% of the words at each of the lower word frequency levels; therefore, due to the above reasons, the 2,000, 3,000, and the AWL levels of the VLT were administered to the participants.

As mentioned above, the VLT results indicated that the participants had good knowledge of the 1,000-word frequency level (*M* > 90%) but had less than 70% knowledge of the lower word frequency levels. Therefore, in order to select experimental target words that would be unknown to the students, but that would still be useful for them in the future, vocabulary items were selected from the 4,500 to 5,000-word frequency level of the Corpus of Contemporary American English (COCA). The COCA was chosen because none of the other teachers in the schools where the research was carried out were aware of it or using it in their classes.

Many vocabulary researchers use nonsense words because, as Webb (2002) pointed out, learners have no prior knowledge of them, they eliminate the need for pretesting, which is time effective, and they eliminate various potential problems concerning the use of pretests. However, Webb also correctly acknowledged that nonsense words have their own potential problems including the fact that they might not represent authentic learning, there could be a carry-over effect after the experiment, and designing nonsense words is a complex process. In addition, there is the ethical problem of whether it is fair to have learners study words that have no real meaning or life application. After taking all of these factors into consideration, low-frequency English words were used, primarily due to ethical considerations; they are suitable for the study because after vocabulary testing the participants and carrying out pilot testing, I am confident that the learners do not know them and will not encounter them during the course of the study. In addition, because they are real words, useful learning can occur; the items might be of benefit to the learners, especially in terms of academic reading in future courses and standardized tests such as the TOEFL iBT and TOEIC. Finally, as they are real words, the treatment better represents the authentic learning of EFL vocabulary.

Because the target words were taken from a list at a lower word frequency level than the participants were studying and they were not included in any of the course materials or activities, nor in the courses that the participants were selected from, or any other courses they were taking, it is unlikely that they met the target words outside of the study. For the participants to have met the words outside of the study, they would have had to carry out significant additional study on their own time, and at a much higher level than the course they were enrolled in. Given my experience and knowledge of these educational contexts and with students enrolled on the courses, this possibility was unlikely.

Obtaining accurate translations (L2 meaning). In order to ensure that the L1 translation of the L2 target items were as suitable as possible, a number of steps were taken. First, initial L1 translations of 80 potential words were independently blind checked by four Japanese-English bilinguals. The initial translations had been made by a bilingual Japanese-English speaking teacher who chose what she felt was the best Japanese translation for each English target item. The list of words and their translations were next sent to four other independent judges: one other bilingual Japanese-English speaking EFL teacher, two bilingual Japanese-English speaking TESOL graduate students, and one bilingual Japanese-English speaking member of staff in the faculty affairs office. Only if consensus was reached by at least three out of the four judges was an item put forward for the final test. Of the original 80 items, 72 (90%) received consensus by at least three out of four judges, four (5%) received consensus by two out of four judges, but none of the judges agreed about the translation of four items (5%). As a result of this process, eight items were removed, and 72 words were selected as possible target items for the experiments described in this and other chapters of this paper.

Pilot testing: Checking the level of difficulty of potential items. The level of difficulty of the remaining 72 vocabulary items was analysed through pilot testing to ensure they would be unknown to the participants in the study. A pilot test was given to 12 students in the same level courses as the participants, but who were taught by other teachers. In addition, the test was administered to six third-year learners who had completed at least one extra year of the English program than the oldest participants enrolled in the present study. These students were at a higher English proficiency level than the participants (TOEIC scores were an average of 200 points higher, and they had demonstrated a sound vocabulary knowledge up to the 3,000 word-level through the VLT). The pilot test first involved students indicating if they had seen the English words before by indicating *yes* or *no*. If the students indicated that they had encountered them previously, they were asked to provide an L1 (Japanese) translation. Because no words were answered correctly by 20% or more of the students, the original 72 candidate vocabulary items remained available for the experiment (Appendix A).

Number of items used in the experiment. In this experiment, all students were required to make their own word cards from information provided to them. It was originally decided that each student would make 30 word cards in one session; however, reconsidering the time constraints within the research setting, and based on well-designed previous deliberate vocabulary studies (Laufer & Rozovski-Roitblat, 2014; Nakata, 2008; Webb, 2002, 2005, 2007), it was decided that participants would make 15 word cards. However, as discussed below, 15 comparison items were added to the test, so the total number of words used in the experiment was 30.

To assess whether any learning takes place a result of taking the three tests, and to take into account other possible confounding variables such as meeting the words outside of the research setting, 15 comparison items were included in the tests. These items were taken from the original sample of words that were selected after the processes described above were carried out.

The rationale for the inclusion of comparison items came from a number of considerations. First, because the three tests—pretest, posttest, and delayed posttest—exposed the learners to both the L1 and L2 versions of the target items as well as the sentences, the participants were repeatedly exposed to the target items during the various testing stages of the study; therefore, it was feasible that some learning might have taken place from the tests. This problem has been noted as a potential flaw of administering pretests (Nation & Webb, 2011, pp. 277; Webb, 2002). Therefore, to assess the possible learning outcomes from taking the tests, comparison items were added. Second, by adding comparison items, a comparison group was not needed, and this allows for a larger number of participants to complete the treatment. This is a major benefit due to only having access to a limited number of potential participants. Furthermore, as noted earlier, the use of control items instead of a control group increases the *ethical validity* of the study.

The only known time that the participants saw these words was while taking the three tests. Therefore, as the participants did not see or study the comparison words during the treatment stage, any knowledge gains made by the participants are attributable to learning from the tests or by being exposed to the words outside of the experiment. If this was the case, these data would need to be used when calculating the learning outcomes of the treatment items.

Selecting a representative sample of vocabulary items. It was deemed important that the test items represented a similar balance of parts of speech as found in natural spoken or written texts in order to increase external and ecological validity (Nation & Webb, 2011, pp. 20–23, 34). According to Johansson and Hofland (1989), English texts are made up of around 50% nouns, 22% verbs, 16% adjectives, and 8% adverbs; therefore, the aim was to match this representation of different parts of speech as closely as possible. Thus, the final 15 treatment items included seven nouns (46.6%), four verbs (26.6%), three adjectives (20%), and one adverb (6.6%). This was also the case for the 15 comparison items. In addition, as far as possible, all target and comparison words were unrelated in either meaning or form; words were chosen that had different meanings from each other, they were not connected topically, and were not similar in spelling.

Further pilot testing to check possible prior knowledge of target words. In order to further check the level of difficulty of the vocabulary items chosen for the experiment, another round of pilot testing was carried out. This testing involved having 21 second-year students from the private woman's university, who were not involved in the study due to logistical reasons, take the pretests that were planned to be used in the study. These students were one year ahead of the first-year participants and at the same level as the second-year participants in terms of their progression through the university English program. No items were known by more than 15% of the participants, and no participant scored more than 25% on any of the pretests. The mean score out of 20 for the active recall test was 0.55, that of the passive recall was 2.50, and the mean on the cloze test was 0.40.

After the pretest scores were calculated, the three students who had scored relatively highly—between 15 and 22%—were interviewed. Interview data highlighted that these students were part of a small percentage of students who had enrolled in an intensive English program called Career English Island (CEI) that is open to second- and third-year students who have passed the application process required to join the program. The students enrolled on this program are generally regarded as being highly motivated and take considerably more English classes in their second, third, and fourth years than the majority of students at the university who take only mandatory English courses. They said that they felt that they might have come across some of the words in the Academic Reading and Discussion course, a class that is designed primarily for second- and third-year students. Given the overall mean scores on the pilot tests, and the fact that even the three students involved in the pilot testing who were enrolled in the CEI program did not score more that 25%, it was decided that the words were at a suitable level; most of the words should have been unknown to the participants.

Selecting standardized information about target items used to make word cards. In terms of presenting the target words and other information that was to be entered onto the word cards, certain decisions needed to be made. The first major decision concerned what information to include on the cards. As noted by Nation and Webb (2011), word cards usually contain the L2 word and its L1 meaning but there are many other types of information that can be put on the cards such as example sentences, part of speech, other word forms, or phonetic information such as main stress patterns (p. 43). However, after (a) reading the literature on word cards, (b) having learners make word cards as a course component in various courses for several years, and (c) informally interviewing students enrolled on these courses, it was decided that only the word meaning (L1 and L2 forms) and example sentences would be included on the cards. One main reason for this was that knowing the meaning of L2 words form and having basic knowledge of word use are two of the most important aspects of word knowledge at the initial learning stages (Nation, 2013, p. 59; Schmitt, 2008).

Therefore, the first information entered on the cards was the L1 target word and their L1 meaning. Although there are many aspects to knowing a word, learning the form-meaning connection is the most basic and vital type of knowledge needed (Nation, 2013, pp. 46–49); therefore, each word card contained a target L2 word and a corresponding L2 translation. As noted above, great care was taken when selecting the L1 meanings of the target words to make any learning outcomes as beneficial as possible.

Second, it was decided that sentences would be added to the word cards. As noted by Nation and Webb (2011), in addition to the L2 and L1 form of the words, learners commonly add an example sentence (p. 43). For Experiment 1, due to theoretical perspectives and empirical support offered by numerous researchers, it was decided that each word card should include an example sentence. The advantages of including example sentences are that additional information is provided by the sentences and the process of adding sentences is potentially beneficial.

The inclusion of example sentences provided the learners with additional useful information about the target items. Adding an example sentence takes relatively little time, and while some studies have suggested that it offers no learning benefit (Barcroft, 2004), it is supported by several other studies, whether it be writing original sentences (Keating, 2008; Liu, 2011; Masson, 2012; Webb, 2007) or copying/reading provided sentences (Pichette, de Serres, & Lafontaine, 2011; Webb, 2005). The main benefit of example sentences is that they provide learners with information about usage of the word instead of just meaning (Nation, 2013, pp. 361, 438). For example, learners can acquire information about how the word interacts grammatically with other words, they might notice collocations, and might better understand typical contexts in which the target words can be used. These types of knowledge come under the *use* component of what is involved in knowing a word (Nation, 2007). As the sentence allows the target item to be met in context and shows the use of the item, it goes beyond developing the form-meaning link, and enhances knowledge of the item by developing other types of word knowledge, something needed to develop mastery of the different word knowledge types (Nation, 2013, pp. 361–365; Schmitt, 2008).

In addition, the process of copying the sentences increases the number of encounters or repetitions with the L2 form of the new word. This repetition occurs because learners need to look at the L2 form of the word again when copying the sentence from the dictionary, and then write it onto their cards. This process provides at least two additional encounters with the L2 target word, while providing access to the L1 meaning. As the number of meetings with a new word has been shown to be important in several studies, this should provide additional learning opportunities (Folse, 2006; Laufer & Rozovski-Roitblat, 2014; Mizumoto, 2013; Newton, 2013). Finally, some researchers have reported that adding sentences is viewed positively by learners and can increase motivation (Al-Murtadha, 2013; Masson, 2012). This is true in my own teaching context where many students have repeatedly stated that they want to add sentences to their cards because they feel they do not learn how to use the words if they just have the L2 form and L1 meaning (Wilkinson, 2017). Increasing motivation is desirable for any language learning activity (Schmitt, 2008), so this is another reason for the inclusion of example sentences within the various experiments outlined in this study.

When creating the example sentences, care was taken to ensure that the length of the various sentences did not vary by more than two words; therefore, all sentences were between 12 and 14 words in length. In addition, to ensure that the sentences were comprehensible, vocabulary analysis was carried out on the example sentences to ensure all non-target words contained in the sentences were high-frequency words; contained in the first 2,000 words of the NGSL.

Another reason for the inclusion of the sentences was that during previous courses and action research projects, many learners had stated that sentences were beneficial to their learning and were adding them to their cards. Therefore, in addition to the pedagogical benefits described above, it was felt that the inclusion of sentences would result in the treatment materials most closely representing what learners do as part of their normal language courses, increasing ecological validity.

Materials: Learner Dictionary and Word Cards

Based on guidelines mentioned in Nation and Webb (2011) to reduce confounding variables and control for internal validity of materials, it was necessary to provide learners with standardized information that they needed to transfer to the word cards during the treatment stage (p. 34, 311). This was done by creating a small dictionary style booklet containing the following information: (a) the target English word, (b) the Japanese translation, and (c) an example sentence in English. The format of the booklet is shown below (see Appendix B for full version):

English word	Japanese word	Sentence
efficient	効率的な	The capital city has a highly ***efficient*** public transport system.

In addition, it was necessary to make sure students were provided with and new how to create word cards**.** Before the study, as part of their normal course requirements the learners had already been engaged in making and studying word cards following the same format described below. Therefore, the participants had been trained in the format that the cards should take and what information must be contained on them. However, directly before the treatment began, the students were given a pro forma and an example to follow as shown below.

Pro forma:

Side A	Side B
English Word	Japanese Word Context sentence (with word represented by *****)

Example:

Side A	Side B
efficient	効率的な The capital city has a highly ********* public transport system.

Instruments

This experiment included three tests in order to better measure depth of word knowledge (Nation & Webb, 2011, pp. 100–104, 306; Schmitt, 2010, pp. 224–225). Numerous researchers including Laufer and Rozovski-Roitblat (2014), Webb (2005, 2007, 2009), and Zhang and Lu (2014) have called for the inclusion of a range of sensitive tests that measure differeing degrees of word knowledge such as the one used in this study. Some researchers such as Webb (2005, 2007, 2009) have included up to ten kinds of tests; however, this was deemed logistically and pedagogically unsound for this study as I did not want learners to spend more time taking the tests than they did carrying out the treatment.

All three tests were marked in a way to give credit for partial knowledge gains and offer a more sensitive assessment of the participants' vocabulary knowledge. In order to do this, a 3-point scoring scale was used based on guidelines given in Barcroft (2004, 2006, 2007a, 2007b) and Nation and Webb (2011, p. 103). Incorrect answers or ones with major spelling errors that would interfere with pronunciation or comprehension, and that were not recognized by a spell checker were scored as 0. Answers that contained a minor spelling error that did not interfere with pronunciation or that was corrected by a spell checker were awarded 0.5 points. A score of 1 was given if the English word was spelled correctly. The answers were scored by two raters who discussed differences of opinions as they arose and came to an agreement about the score awarded.

The pretests, immediate posttests, and delayed posttests were the same in terms of format, vocabulary items, and sentences; the only difference was the order that the items were presented. The order of the vocabulary items and sentences were changed for each test in order to reduce the chance of list effects or serial learning. The L1 and L2 items and sentences were the same as the information provided in the dictionary style booklet described above.

A brief description of each of the three test types is provided below (see Appendix C to E for a full copy of the pretest):

Active recall test. The active recall test requires the participants to provide the English translation of a given Japanese word as shown below.

Japanese word	English word
効率的な	efficient
受ける	undergo

According to Laufer and Goldstein (2004) and Laufer and Rozovski-Roitblat (2014), this type of test requires active recall of the L2 target word form, and for this reason is more difficult than passive recall tests and multiple-choice tests. This test was chosen as it represents the type of knowledge learners need when recalling and using L2 words in speaking or writing activities (Laufer & Goldstein, 2004).

Passive recall test. The passive recall test requires participants to provide the L1 translation of a given L2 word as shown below.

English word	Japanese word
efficient	効率的な
ancestor	祖先

Laufer and Rozovski-Roitblat (2014) noted that this type of test requires passive recall of the L2 target word form, and while this is easier than the active recall tests, it is more difficult than a multiple-choice test. This test was chosen to offer an additional measure of vocabulary knowledge, and because this type of test better represents the most lifelike knowledge learners need, as it is the type of knowledge needed to read L2 texts or listen to L2 speakers (Laufer & Goldstein, 2004; Laufer & Rozovski-Roitblat, 2014).

L2 cloze test. This test required participants to enter the correct English word in a gap in a sentence. Between one and three initial letters were provided to make the test more manageable given the short amount of study time and to limit correct answers to only those provided during the treatment. The number of letters was based on how many letters were deemed necessary to ensure that there was only one possible correct answer (Schmitt, 2010, p. 231). The sentences containing the target items in the test were the same as the sentences that students copied on to their word cards during the treatment. When answering the test, the items could have been answered by using the context, the first letter, or both, but learners were instructed that only words included in the word card treatment could be used. A sample item is as follows:

Many people agree that it is a good idea to **pr__________** smoking on campus.

According to Nation and Webb (2011, pp. 104–114), Schmitt (2000, pp. 152-154), and Schmitt (2010, pp. 229–231), this type of activity assesses learners' knowledge of the use of a particular word and might involve issues such as knowledge of grammatical functions and collocations. However, given that the sentences in the test we the same as on the word cards, features of use are provided, and no generative use is required, the students might have remembered the sentences from the cards and the cloze test can be seen as another form of active recall test. At best, the cloze test would only measure knowledge of how to use the target words to a very limited extent.

Piloting

Once the final words had been selected and compiled into the tests, additional piloting was conducted to further check the target item difficulty, assess the time required for the tests, and make sure that the tests were easily understandable. The piloting was conducted with 20 students from a class who were deemed to have similar English language proficiency as the participants.

First, the participants took the pretest that was to be used in Experiment 1. The students were observed while the test was being administered, and some were asked if there were any issues concerning the test design or instructions. In order to investigate how long was needed for the treatment, pilot testing was again carried out. The dictionary style vocabulary handout and 20 blank word cards were given to the 19 students involved in the piloting. These students were instructed in how to make the cards and were then observed and timed by me, and when a particular student finished making the cards, the completion time was noted. The minimum time recorded was 12 minutes and 56 seconds, and the maximum was 15 minutes and 9 seconds. The mean time was 14 minutes and 24 seconds. Therefore, it was decided that participants would be given one minute to make each word card for a total of 15 minutes.

Overall, the results of the pilot testing indicated that (a) most target words should be unknown by the participants, (b) the tests were easily understood and completed correctly by all participants, and (c) 20 minutes was needed to complete all three tests.

Procedures

All tests and treatments were administered in the normal classroom setting. In addition, the participants were not informed of the existence of any of the tests or treatments in advance.

Careful consideration was given to the order the tests were administered in because exposing participants to information about target items on the tests can affect results on subsequent tests (Laufer & Goldstein, 2004; Liu, 2011). The cloze test was administered first as it reveals little to no information about the target items. The second test was the active recall test because this test only displays the L1 translation of the target L2 words. Finally, the participants took the passive recall test; it is likely that the scores on this test were positively affected by the two preceding tests (Laufer & Goldstein, 2004).

After the cloze test was completed, tests were collected from all students before the active recall test was handed out. Likewise, the active recall tests were collected before administering the passive recall test.

At each of the three testing stages, the participants were given as long as they needed to take each of the three tests. Each of the three pretests (cloze, active recall, and passive recall) took between 4 and 5 minutes to complete (15 minutes in total), the immediate posttests took from 8 to 11 minutes each, and delayed posttests required 7 to 11 minutes to complete.

Pretests. The three tests were administered as pretests two weeks before the treatment. This time delay was based on that used in numerous previous studies, as well as advice in the vocabulary learning and testing literature that indicates that two weeks provides sufficient time for learners to forget the contents of the tests and not affect the outcomes of the treatment (Nation & Webb, 2011, pp. 276–278). As mentioned above, the participants were given as long as they needed to take the tests and were asked to wait until the test was complete

Treatment and posttests. In order to answer research question 1, on the day of the treatment, the participants were given 15 minutes to make 15 word cards before taking an immediate posttest. First, cards were created using the standardized information in the dictionary style booklet and following the format described earlier. Students were given the dictionary handout and blank word cards and told they had 15 minutes to make the cards. The only other materials students were allowed was a pencil and eraser or a pen. They were instructed not to study from the cards if they finished early (see Analysis of Quantitative Data regarding this point). Immediately after the treatment the cards were collected and retained, then the participants then took the immediate posttest. During the treatment, the participants were observed to ensure that only those who fully completed the treatment by making all 15 word cards were included in the analysis.

Next, in order to answer research question 2, delayed posttests were administered two weeks after the treatment. These data were used to calculate the amount of attrition that takes place if no additional deliberate study is carried out. Two weeks was chosen based on guidelines provided by Nation and Webb (2011, pp. 278-279) and the results of studies showing that significant attrition in knowledge occurs rapidly if learners do not encounter or study vocabulary items for a relatively short period of time, be it a few days or a few weeks (Brown, 2012; Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014; Waring & Takaki, 2003).

Quantitative Data Analysis

Quantitative data for Experiment 1 were analyzed using one-way repeated-measures MANOVAs and one-way repeated-measures ANOVAs using IBM SPSS statistics version 21. Repeated-measures analysis was chosen because of the within-subject design and the inclusion of multiple test stages. In this within-subjects design, all participants experienced the same treatments, and in the pretest-posttest design, all participants and were tested over three time points. This means that the cases observed in the first observation (pretest scores) are directly linked with the cases in all other observations (posttests), and a repeated measures analysis takes this into account (Field, 2018, pp. 651–653). Pairing measured data points is done when it is difficult to control all the factors that could influence the results and ensures that such factors are equally likely affect each data point (Verma, 2016). Examples of such uncontrollable factors include participants' motivation levels, feelings about the treatment, or aversion to taking tests.

One-way repeated-measures MANOVA analysis. Pretests data were examined using one-way repeated measures MANOVA. The within-subjects factor was treatment with two levels (word card treatment and comparison), and there were three dependent vocabulary pretest measures: passive recall, active recall, and cloze test.

However, before running and interpreting the MANOVA results, analysis was carried out to confirm that the assumptions of repeated-measures MANOVA (Field, 2018, p.651; Laerd Statistics, 2018; Verma, 2016) were met. First, there was one independent variable with two categorical, related groups (treatment and comparison), and three dependent variables measured on an interval scale of between zero and 30. Second, participants were measured at all time points and received all treatments. Next the presence of univariate or multivariate outliers was checked in SPSS using boxplots and a measure called Mahalanobis distance. The boxplots showed only six univariate outliers, and after comparing the Mahalanobis Distances to a chi-square distribution, five of these participants had a probability of less than .001 so were considered multivariate outliers. Removing these participants was not ideal as it would reduce the sample size and lead to a loss in statistical power (Field, 2018, p. 753), so the effect of the outliers was determined by examining the original MANOVA results with that of analysis with outliers removed. This analysis revealed that the presence of these outliers did not affect the results of the MANOVA analysis, indicating the outliers could remain in sample population. Next, the linear relationship between each pair of dependent variables for all combinations of groups on the two independent variables was confirmed in SPSS by plotting a scatterplot matrix for all combinations of group on the two independent variables. The scatterplot matrix showed that there was a weak linear relationship between the pretest and posttests, and a strong linear relationship between the immediate and delayed posttests on all dependent measures for both treatments. Next, the normality of each of the dependent variables for each of the related independent variable groups was tested using the Shapiro-Wilk test of normality and Normal Q-Q plots. The Shapiro-Wilk test was significant for many of the combinations, indicating that the normality assumption was violated, however, the Q-Q plots showed that the normality did not stray too far from the expected values. Further analysis of the data indicated that z-skewness was > 3.29 for all tests but z-kurtosis was < 3.29 for most tests. Blanca, Alarcón, & Arnau (2017) concluded that non-normality is not an important issue where ANOVA and MANOVA analyses are concerned: "regardless of the degree of deviation from a normal distribution, sample size, balanced or unbalanced cells, and equal or unequal distribution in the groups" (p. 554). Also, Khan and Raynor (2003) concluded if there is a problem with non-normality, it concerns kurtosis not skewness. Taking the above into consideration, it was decided that the violation was not too severe. However, as an extra precautionary step, during the analysis stage, Friedman and Wilcoxon non-parametric tests were carried out and results confirmed the findings of the MANOVA analysis. Finally, a Pearson correlation analysis showed that the dependent variables were mildly correlated and there was no multicollinearity. As stated by Laerd Statistics (2019), running MANOVA analysis is not recommended if correlations are too low, but especially if there is multicollinearity; “if the correlations are too high (greater than 0.9), you could have multicollinearity … this is problematic for MANOVA ” (para. 16). Five of the pretest pairs were correlated between r = .15 and .17, four were correlated between r = .20 and .28, and four were correlated between r = 30 and .55.

One-way repeated-measures ANOVA analysis. Next, various analyses of the comparison and treatment groups’ mean scores on the dependent measures over the three time points were carried out using a series of one-way repeated-measures ANOVA. This analysis was chosen over a repeated-measures MANOVA after Pearson correlation analysis indicated that many of the immediate and delayed dependent posttest measures were not sufficiently correlated. For example, the active recall pretest and delayed posttests were correlated at r = -.019, the passive recall pretest and posttest were correlated at r = .195, and the pretest cloze means were not correlated to the delayed posttest cloze means (r = -.03,), and the active recall, passive recall, and cloze delayed posttests were all correlated at below r = .20. While there is no concrete cut point, Laerd Statistics (2019) states that repeated-measures ANOVA is recommended over repeated-measures MANOVA if correlations are below r = .20. Based on these results, a major assumption of repeated-measures MANOVA analysis was violated, so one-way repeated-measures ANOVA analysis was carried out.

As described below, the assumptions of repeated-measures ANOVA (Field, 2018, pp. 663, 753; Laerd Statistics, 2018; Verma, 2016) were checked before interpreting the

ANOVA results. First, assumptions about the data types were met as the independent condition variable had two related groups (word cards and comparison), the dependent vocabulary measures were scored on an interval ration from zero to 30, and all participants underwent both conditions and were tested at three time points. Next, there were no significant outliers in the related groups. Next, normality checks were carried out and while the Shapiro-Wilk test was significant for many of the combinations, the Q-Q plots showed that the normality did not stray far from the expected values. As noted earlier, Blanca, et al. (2017) and Khan and Raynor (2003) concluded that non-normality is not an important issue where ANOVAs are concerned, so the violation was not such a significant problem. In addition, non-parametric Friedman test results indicated that the ANOVA results were not affected as a result of breaches in normality. Finally, sphericity, which assumes that the variation within conditions is similar and that no two conditions are any more dependent than any other two, was checked using Mauchly's test of sphericity, and as described in the analysis section, if significant the Greenhouse-Geisser correction was used (Laerd Statistics, 2018).

Qualitative Data Analysis

In addition, qualitative data were collected via observations of all students present in each class during each stage of the experiment. In addition, informal semi-structured interviews were carried out with two students from each intact class (n = 8). The observation data provide insights into what the students did when making the vocabulary cards during the cross-sectional treatment. The informal interview data offer further insights into the students' perceptions of this type of controlled word card method (see Appendix F for the interview protocol). The qualitative data were analysed and coded in order to offers holistic and triangulated data about some of the individual learner variables that might have affected the quantitative learning outcomes (Hatch, 2002, pp. 93–133; Riazi & Candlin, 2014). This was done through content analysis with the bulk of observations and responses being coded and placed under pre-structured themes and sub-headings that arose from the research questions and the related literature concerning these issues, for example, *General Perceptions*, *Time/Speed*, and *Convenience/inconvenience*. However, during the analysis some unplanned or unexpected themes emerged, so a number of additional codes, themes and sub-headings, were added, for example, *Word-selection,* and *Simplicity*. (Hatch, 2002, pp. 155–162).

Results: Learning Outcomes of Making Word Cards

Analysis of the Pretest Data

The first analysis involved running a one-way repeated-measures MANOVA on the pretest means. The within-subjects factor was treatment with two levels (word card treatment and comparison), and the three dependent vocabulary measures were passive recall, active recall, and cloze test. This analysis was carried out to investigate if there were any statistically significant differences between the pretest means for the treatment items and comparison items on the three dependent measures at the start of the study.

As determined by the repeated-measures MANOVA, there was no statistically significant main effect for treatment at the pretest stage, Wilks $\Lambda = .93$, $F(3, 69) = 1.77$, p

= .16, partial η^2 = .07, observed power = .44; thus, there was no statistically significant difference between the treatment and control mean scores on the dependent measures before the treatment began. These results indicate that there were no significant differences between participants knowledge of the target and comparison items at the beginning of the study, indicating that the words were of similar difficulty or familiarity level. The descriptive statistics shown in Table 7 provide further information regarding the pretest data.

Table 7. *Descriptive Statistics for the Three Dependent Pretest Measures*

	Treatment	Comparison
Pretest Active Recall		
M	.13	.10
SD	.41	.42
SE	.05	.05
95% CI	[.03, .22]	[.00, .20]
Pretest Passive Recall		
M	.53	.43
SD	.89	.80
SE	.10	.09
95% CI	[.32, .74]	[.24, .62]
Pretest Cloze		
M	.17	.03
SD	.56	.24
SE	.07	.03
95% CI	[.04, .30]	[- .03, .08]

Note. *N* = 72. CI = Confidence Interval. There were 15 target items and 15 comparison items.

As shown in Table 7, the mean scores were well below 1.00 on all measures, meaning that most participants knew none of the words selected for use in the experiment at the pretest stage. The descriptive statistics showed that the minimum score for the treatment items and comparison items was 0 and the maximum was 2. Taken together with the means, most participants did not know any of the words, and no participants new more than one treatment or comparison item (The partial scoring method meant that each word had a maximum score of 2). Given that there were no statistically significant differences between the target and comparison item means scores at the beginning of the study, and the words were largely unknown, analysis of the comparison items over the three time points will indicate how likely it is that the target items were learned from the repeated testing stages or via exposure outside of the experiment.

Analysis of the Comparison Items

Next, in order to investigate if there were any significant changes in the comparison item mean scores over time, analysis was carried out using two separate one-way repeated-measures ANOVAs. The factor for each repeated-measures ANOVA was time of testing (three levels: pretest, immediate posttest, and delayed posttest) and the dependent variable was one comparison item knowledge measure. For the first repeated-measures ANOVA, the dependent measure was the active recall test, for the second it was the passive recall test, and for the last it was the cloze test.

The repeated-measures ANOVAs indicated that no statistically significant learning took place over time on any of the dependent measures of comparison item knowledge. First, there was no main effect for time for the active recall measure of the comparison items, Wilks $\Lambda = .94$, $F(2, 142) = 1.96$, $p = .14$, partial $\eta^2 = .03$, observed power = .40. Second, a repeated measures ANOVA with a Greenhouse-Geisser correction determined that there were no statistically significant differences in the mean passive recall scores for the comparison items between time points, $F(1.73, 122.63) = .34$, $p = .71$, partial $\eta^2 =$

.01, observed power = .10. Greenhouse-Geisser correction was selected because the Mauchly's sphericity test indicated that the sphericity assumption was violated. Finally, there was no main effect for time on the cloze test means as determined with a Greenhouse-Geisser correction, $F(1.65, 117.33) = 1.95$, $p = .15$, partial $\eta^2 = .03$, observed power = .36.

The above results show that there were no significant differences among the mean scores on the comparison item measures between the pretest and delayed posttest stages, meaning that no learning took place. The lack of significant difference between the comparison item means over the three time points indicate that the participants did not learn from the test and were not exposed to the target words outside of the experiment. Therefore, this analysis eliminated the two major possible confounding variables that could not be controlled in the design of this study, and as a result, it can be safely assumed that any learning that took place under the treatment condition (word card making) can be attributed to only the treatment.

Analysis of the Treatment Items

In order to investigate if there was (a) significant vocabulary learning as a result of making word cards, and (b) significant attrition in word knowledge after not studying for a two-week period, an analysis was carried out using a series of one-way repeated-measures ANOVAs. As mentioned, this analysis was chosen over a repeated-measures MANOVA because the Pearson correlation showed that not all the dependent measures were sufficiently correlated.

The within-subject factor for each ANOVA was time of testing (pretest, immediate posttest, and delayed posttest), and the dependent variable was one of the measures of vocabulary knowledge: active recall, passive recall, and cloze test. Statistically significant effects for time were found on all three of the dependent vocabulary measures. First, a one-way repeated-measures ANOVA determined that the mean scores on the active recall measure of vocabulary knowledge differed significantly between the time points, $F(2, 142) = 241.11$, $p < .001$, partial $\eta^2 = .77$, observed power = 1.00. Second, a repeated-measures ANOVA with a Greenhouse-Geisser correction indicated a main effect for time on the passive recall test mean score; $F(1.74, 123.46) = 665.55$, $p < .001$, partial $\eta^2 = .90$, observed power = 1.00. Finally, there was a main effect for time on the cloze test means (Greenhouse-Geisser adjusted); $F(1.65, 117.31) = 365.61$, $p < .001$, partial $\eta^2 = .84$, observed power = 1.00. These results show that there was a statistically significant effect of time of testing for the word card treatment on all dependent measures; thus, vocabulary learning and/or attrition took place between the pretest, posttest, and delayed posttest stages as a result of making word cards. In order to begin to understand the differences, it is useful to look at the descriptive statistics for each dependent measure at the three time points (see Table 8).

As shown in Table 8, the mean scores on all dependent measures increased from pretest to posttest, but then decreased from posttest to delayed posttest (2 weeks later). However, the means at the delayed posttest were still considerably higher than at pretest. This result shows that while learning took place from making the word cards, after two weeks of not studying or seeing the words, considerable attrition occurred, but there was still some long-term learning as measured by the three vocabulary knowledge measures at delayed posttest.

Table 8. *Descriptive Statistics for the Pretest, Posttest, and Delayed Posttest Means*

	Active	Passive	Cloze
Pretest			
M	.13	.53	.17
SD	.41	.89	.56
SE	.05	.10	.07
95% CI	[.03, .22]	[.32, .74]	[.04, .29]
Posttest			
M	18.06	21.44	15.67
SD	7.40	5.96	5.84
SE	.87	.70	.69
95% CI	[16.32 .19.76]	[20.04, 22.84]	[14.30, 17.04]
Delayed Posttest			
M	9.31	16.89	7.31
SD	6.83	5.35	4.26
SE	.80	.63	.50
95% CI	[7.70, 10.91]	[15.63, 18.15]	[6.30, 8.30]

Note. N = 72. CI = Confidence Interval. 15 target items: maximum possible score is 30.

Active recall test results. To investigate which of these changes that took place over time on the active recall measure were statistically significant, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at < .017). The pairwise comparisons data are displayed in Table 9.

Table 9. *Pairwise Comparisons for the Active Recall Test Over Three Time Points*

	Posttest to pretest	Delayed posttest to pretest	Delayed posttest to posttest
Mean difference	17.93	9.18	-8.75
95% CI	[15.82, 20.04]	[7,20, 11.16]	[-10.66, -6.84]
SE	.86	.81	.79
p-value	< .001	< .001	< .001

Note. N = 72. CI = Confidence interval. 15 target items: maximum possible difference is 30.

As seen from the pairwise comparisons for the active recall test over the three time points, there were significant vocabulary gains between the pretest and posttest and pretest to delayed posttest, but significant attrition also occurred from posttest to delayed posttest. First, the mean gain from pretest to posttest was 17.93 ($p < .001$). As the partial scoring method meant that a correct answer is awarded a score of 2 points, the maximum score was 30 (partially correct answers were awarded 1 point). Therefore, the mean gain of 17.93 is equivalent to 8.96 words out of 15 (60%). Next, the mean gain from pretest to delayed posttest was also significant ($p < .001$), with a mean increase of 9.18 (4.59 words; 30%). Therefore, while the treatment resulted in semi-long-term memory of some target words, there was a substantial amount of word knowledge attrition. The attrition from posttest to delayed posttest was -8.75 (4.37 words), meaning approximately 49% of posttest active recall knowledge was lost by the delayed posttest.

Passive recall test results. Next, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at $< .017$). Table 10 shows the pairwise comparison data for the passive recall test means over the three time points. There were significant vocabulary gains between the pretest and posttest with a mean increase of 20.92, which is equivalent to 10.46 words out of 15 (70%). Second, significant gains were also observed between the pretest and delayed posttest with a mean gain of 16.36 (8.18 words; 54%). However, there was again statistically significant attrition from posttest to delayed posttest (-4.56 = 2.28 words); approximately 22% of previous passive recall knowledge was forgotten.

Table 10. *Pairwise Comparisons for the Passive Recall Test over Three Time Points*

	Posttest to pretest	Delayed posttest to pretest	Delayed posttest to posttest
Mean difference	20.92	16.36	-4.56
95% CI	[19.23, 22.60]	[14.84, 17.88]	[-5.74, -3.37]
SE	.69	.62	.48
p-value	< .001	< .001	< .001

Note. N = 72. CI = Confidence interval. 15 target items: maximum possible difference is 30.

The mean gains for the passive recall test were higher than those of the active recall test; this result was expected because the passive recall test, which involved providing the Japanese translation for a given English target word, is the easier test (Nation & Webb, 2011, p. 100). However, it is also clear that less passive knowledge attrition occurred; 49% of the words tested actively were forgotten compared to only 22% of the words on the passive test. Therefore, passive knowledge gains were higher than active knowledge and were better retained over time when no further study was carried out after initially making word cards.

Cloze test results. Next, to identify where any significant differences for the cloze test exist over the three time points, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at < .017). Table 11 shows the pairwise comparison data for the cloze test over the three time points and as seen, there were statistically significant vocabulary gains between the cloze pretest and posttest (p < .001). The mean gain was 15.50, which is equivalent to 7.75 out of 15 words (52%). There were also statistically significant gains between the pretest and delayed posttest with a mean gain of 7.14 (3.57 words; 24%), but there was also statistically significant attrition from the posttest to the delayed posttest (-8.36 = 4.18 words: 54% attrition from posttest to delayed posttest).

Table 11. Pairwise Comparisons for Mean Differences on the Cloze Test Over Three Time Points

	Posttest to Pretest	Delayed Posttest to Pretest	Delayed Posttest to Posttest
Mean difference	15.50	7.14	-8.36
95% CI	[13,80, 17.19]	[5.87, 8.40]	[-9.57, -7.15]
SE	.69	.52	.49
p-value	< .001	< .001	< .001

Note. *N* = 72. CI = Confidence interval. 15 target items: maximum possible difference is 30.

When comparing these results with the descriptive statistics provided in Table 8, the mean gain scores on the cloze test were lower than those on the active and passive recall tests. In addition, the attrition rate from the posttest to the delayed posttest was higher for the cloze measure of vocabulary knowledge (see Table 12).

Table 12. Levels of Attrition From the Posttest to the Delayed Posttest on the Three Dependent Measures

	Active recall	Passive recall	Cloze
Mean difference	-8.75	-4.56	-8.36
% change	49.00%	22.00%	54.00%

Note. *N* = 72. 15 target items: maximum possible mean difference is -30.

The kind of knowledge needed for the cloze test falls under what Nation (2013) has referred to as the *Use* component of word knowledge; knowing how or where to use a word in a given context (p. 49). The above results suggest that this type of knowledge is difficult to develop by making word cards, even after copying sentences onto the cards, and is forgotten more rapidly than other types of word knowledge. The post-treatment task was simple because it required students to place those words in the same sentences as opposed to writing original sentences, but even this easier task proved difficult after making the word cards.

Summary of Results of Quantitative Data Analysis

The main results of this experiment were that the process of making word cards resulted in significant short-term and long-term vocabulary learning. The learning outcome results can be interpreted with confidence as no learning of the comparison items took place, so it is unlikely that participants were exposed to the words outside of the experiment and the positive learning outcomes can be attributed to the word card making treatment. The results indicate that spending relatively little time making word cards, approximately one-minute per word, produced significant gains on all three types of vocabulary knowledge, but passive recall knowledge was developed the most, followed by active recall knowledge. However, as indicated by the cloze test results, only limited knowledge of how to use the words in context was developed even when example sentences were provided. In fact, as mentioned, the cloze test can be seen as another form of active recall because features of use are provided, and no generative use is required, indicating that word cards are not an effective method of building productive knowledge.

First, the results of the analysis of the treatment outcomes show that significant short-term memory of the target vocabulary was developed as a result of making word cards in a 15-minute period. Immediately after making the cards, on average, participants were able to passively recall 10.5 of the 15 target words (70%), actively recall 9 words

(60%), and correctly place 7.5 of the target items (52%) in a sentence. These results also show that while participants made gains in all knowledge types, making word cards led to better short-term passive recall knowledge than active recall knowledge or knowledge of how to use the words in a given context, the latter of which was the least developed knowledge type of the three.

Second, making word cards also resulted in longer-term vocabulary knowledge gains, but significant attrition also occurred during the two-week period between the posttest and delayed posttest. Two weeks after making the cards, with no additional study or exposure to the target items, on average students were able to actively recall 4.65 target words (31%), passively recall around 8.45 words (56%), and correctly place 3.65 words (24%) in sentences. Given that the total study time was only 15 minutes, and two weeks had passed, these small gains are still impressive; however, comparing the immediate posttest and delayed posttest means show a statistically significant decline in knowledge on all dependent vocabulary measures; however, passive recall knowledge gained from making the word cards was more robust. Only 22% of previously learned passive recall knowledge was forgotten, but active recall ability declined by almost 50% and the ability to place the words in a sentence declined by 54%. In summary, these results show that while the word card making process produced significant long-term memory of the target vocabulary, if the cards are not studied soon after they are created, much of the new knowledge is forgotten, especially active recall knowledge and knowledge of how to use the words in context.

Results of Qualitative Data Analysis: Observations and Interviews

The observation and interview data served two main purposes in this study. The data provide results that support the research design and choice of treatment, but also add useful information about learner factors that might have affected the statistical outcomes reported above.

First, the qualitative data add information that increases the internal validity of the research design and subsequent results. The observational data indicate that enough time was given for participants to make the word cards in the way they were supposed to, and all participants produced the required number of cards. During the treatment, none of the participants had any problems understanding or following the instructions. In addition, almost all students were easily able to complete the treatment (make the desired number of word cards) in the short time given and did not carry out any additional study after making the word cards. Only three students from the final 72 participants (4%) did not fully complete the task on time; they did not have time to add the sentence to the last word card. However, they were instructed to quickly finish the last card as I was collecting the study materials from the other participants. In addition, seven students finished the treatment approximately one minute early, but none of them chose to study the words; they either just relaxed or did homework activities. When asked about the time given, all eight students interviewed stated that there was enough time to make the cards and they did not feel rushed; they confirmed that there was either no or very little time to study the cards. Four of the students interviewed were those who had finished one or two minutes early, but they said as they had been instructed not to study the words if they finished early, they did not do so.

Second, the qualitative data shed light on how the controlled treatment compares to a more natural word card making process. First, observations supported the treatment design because the students were engaged in the typical processes and actions involved in natural word card making. Students first looked at L2 word presented in the learner dictionary and then checked their L1 meaning. When transferring the information to the cards, most students appeared to consult the dictionary multiple times. For example, when copying the L2 form, most students looked at the dictionary two times to ensure correct spelling, and most also looked at the dictionary twice while writing the L1 meaning. They then look at the dictionary again to read the example sentence and referred to it multiple times during the process of copying the sentences. Some students read the sentence out load, focusing on the pronunciation of the target word. In addition, nine students (12.5%) asked me to demonstrate the pronunciation of the word after the experiment had finished. Second, when interviewed about how the controlled word card making process compared to the way they usually make their word cards as part of the course requirements, participants stated that it was very similar. The students confirmed that they had to look at the dictionary multiple times to complete each word card. However, they also stated that when using their own dictionaries, there is often multiple L1 choices for each L2 word or vice versa, and they preferred being able to choose the meaning of the word themselves. On the other hand, four of the students mentioned that they thought the quality of the example sentences was much better than those provided in their own dictionaries. They stated that their electronic dictionaries often give only phrases, not sentences, but the sentences provided during the treatment were longer and provided more context. Overall, even though the treatment was not a natural word card making task, it involved the very similar processes and levels of engagement, and students thought it was very similar to what they normally do, but some did state that they preferred having choice of L1 meaning of an L2 target word.

The qualitative data also highlighted affective factors relating to making word cards. First, the interview data showed that students have a positive view of making word cards for deliberate vocabulary study. All students stated that word cards are a quick and easy method of making materials that can be later used for deliberate vocabulary study. Students stated that the process was simple and easy to do, and that writing the words themselves would help them to remember them in the future. In addition, the interviewees stated that they liked the two-sided nature of word cards because it allows for more active and passive recall than a vocabulary notebook where the L1 and L2 forms are placed side-by-side. For example, one participant stated that “When I make the cards, when writing side two, I think about what is on side one.” Overall, the interviewees had a positive view of making word cards, stating that it was an easy process and it forced them to look at and think about the target words multiple times.

However, when asked about the value of adding example sentences to the cards, the results were mixed. Half of the students said that the sentences were helpful, and they liked adding them to their cards during normal study practice, but half thought it took too long and was not so helpful. Those who liked adding sentences thought that it forced them into writing the target word again (more attention to form), helped them get a better idea about the meaning of the word, for example, its similarity to or difference from the L1 form (attention to meaning), and it allowed them to see how it can be used (attention to use or grammar). However, comments by those who were against adding sentences included that the added time made the process laborious, the sentences did not offer any useful extra information about meaning (they got what they needed from the L1 form), and unless they were going to memorize the sentence, then they did not see the point of adding sentences.

A further finding was that, based on the observation and interview data, making word cards results in multiple meetings with the target words. Students stated that when making the cards, (a) they needed to look at the target word multiple times, and (b) the two-sided nature increased the number of times they needed to look at the words. Most students stated that they first looked at the L2 target word and then checked their L1 meaning. They then wrote the L1 word (sometimes without checking it again) before looking at the L2 form again and copying it onto the cards. Four of the students independently stated that they usually had to look at the L2 form more than once when copying it because they paid particular attention to the spelling. When asked about this, two of the remaining four interviewees agreed that they often needed to check the L2 word multiple times when copying it onto their cards, especially if it was a longer word. They then needed to look at the L2 target item at least one more time, but often two or three times, when writing the sentence. Based on the observation and interview data described above, all students met the L2 target word at least five times during the word card making treatment used in this study, but most students checked spelling multiple more times, increasing the number of meetings with the word. The basic process students underwent, and number of meetings is as follows:

1. Look at the L2 form then L1 meaning (Meeting 1)
2. Write the L1 form on side A
3. Look at L2 form again (Meeting 2)
4. Write the L2 form and check spelling on completion (Meeting 3)
5. Read sentence (Meeting 4)
6. Copy sentence (Meeting 5)

As seen above, the process of making word cards involved multiple meetings with the target words (repetition), which is viewed as essential to vocabulary learning, especially in terms of transferring new knowledge to short-term memory (Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014).

In summary, students view the process of making word cards favourably and are aware of the benefits of word cards noted in the literature. In addition, making word cards involves at least five meetings with each L2 target word, something that likely contributes to the learning outcomes. Furthermore, the observation and interview data support the research design and suggest that the learning outcomes observed would be similar for a natural word card making study where students have free choice of L1 meaning and sentence selection.

Discussion

The results of this study not only highlight a number of key points that contribute to the existing vocabulary literature, but also offer valuable new findings and suggestions that are relevant to both researchers and teachers. For example, some findings are supportive of previous claims in the literature on word cards, and in a broader sense to deliberate learning techniques in general, others call into question conclusions previously made, but the main results add completely new information to the word card literature.

First, the results for research question 1 and 2 offer the most unique contribution to the word card literature as they isolate the short and semi-long-term learning effects of simply making word cards. This is the first study to carefully control a word card treatment so as to be able to accurately assess the effects of the word card making process, without further study being carried out. Therefore, the positive learning outcomes found in the study provide research-based support for statements such as “We know from experience that after making the cards, learners can recall from 30-70% of the meanings on the first run through the cards” (Nation & Webb, 2011, pp. 42, 102) and “By writing the cards the learner pays more attention to the word form of the foreign-language word than is usually the case … which has a positive effect on the productive retention” (Mondria & Mondria-de Vries, 1994, p. 53). Further details about specific results and their implications are discussed below.

The results for research question 1 suggest that the process of making word cards produces statistically and pedagogically significant short-term vocabulary learning outcomes. Immediately after making the cards, on average, participants were able to recall between 70% and 50% of the 15 target words. These results are supportive of the literature citing the efficacy and importance of deliberate vocabulary learning in general (Carter & McCarthy, 2014, p. 12; Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014; Nation, 2013, pp. 2–3; Schmitt, 2008), and the opinion that word cards are an effective deliberate vocabulary learning technique (Ballance & Cobb, 2020, pp. 324-325; Mondria & Mondria-de Vries, 1994; Nation, 2013, p. 437). More specifically, the results contribute to the growing body of empirical evidence showing that word cards are a quick and effective deliberate study technique (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014). In addition, given the relatively simple but repetitive nature of the word card making task, the immediate posttest results lend support to the levels of processing view that even shallow processing can help learners commit new vocabulary knowledge to short-term memory (Craik & Lockhart, 1972), especially if the initial task offers the chance for repetition (Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014; Nation, 2013, pp. 107–110).

Next, the main findings for research question 2 are that (a) some of the new vocabulary knowledge gained from making word cards is also committed to long-term memory, but (b) much of the initial knowledge is forgotten (does not enter long-term memory) if no further study is carried out within a relatively short period of time. Two weeks after making the cards (with no additional study or exposure to the target items), students demonstrated knowledge of between 30% and 55% of the target words, which is a positive and pedagogically significant result given the short time-on-task. However, a comparison of the immediate and delayed posttest results showed that participants had lost between 22% and 54% of previously learned vocabulary knowledge. These findings support the assertion that repeated study is vital if vocabulary knowledge is to be committed to longer-term memory (Nation, 2013, pp. 107–110; Schmitt, 2008) and are similar to previous research findings (Kuo & Ho, 2012; Oberg, 2011; Nakata, 2008; Waring, 1997, 2004). Moreover, the findings support the view that while tasks involving only relatively shallow processing (Craik & Lockhart, 1972) and low task-induced involvement load (Laufer & Hulstijn, 2001) can be effective for transferring new vocabulary knowledge into short-term memory, they are not particularly effective for improving long-term memory, echoing the findings of other vocabulary researchers (Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014; Watanabe, 1997).

Considering the findings for research questions 1 and 2 together, another significant contribution that this study makes is that the word card making process results in better short and long-term passive recall knowledge than active or productive knowledge. Not only were the passive recall mean scores higher on both the immediate and delayed posttests, but the passive recall knowledge was much less sensitive to attrition over time. These findings support previous research stating that it is easier to build receptive vocabulary knowledge than productive knowledge (Komachali & Khodareza, 2012; Kuo & Ho, 2012; Waring, 1997; Webb, 2005, 2007, 2009). For a more detailed discussion of the literature on this issue, see Behzadi and Haji Pour Nezhad (2014), and Webb (2002). This finding was made possible as this study goes beyond that of many other word card studies (Altiner, 2011; Emrah & Omur, 2010; Hung, 2015; Nikoopour & Kazemi, 2014) in that it uses three vocabulary learning outcome measures (active recall, passive recall, and cloze test) and a partial scoring method, all of which allowed for a more accurate assessment of learning outcomes (Barcroft, 2007a, 2007b; Laufer & Goldstein, 2004; Laufer & Rozovski-Roitblat, 2014; Webb, 2007, 2009).

Next, while the specific learning effect of using example sentences in the word card making treatment is unknown, a great deal of support for such an activity can be found in the literature (Al-Murtadha, 2014; Behzadi, & Haji Pour Nezhad, 2014; Folse, 2006; Nation, 2013, p. 438). For example, Folse (2006) found that even simple fill-in-the-blank exercises produced positive learning outcomes, and Webb (2005) found that writing original sentences was beneficial. There are a number of specific task-induced features of the sentence writing task used in the treatment that are likely to have increased the learning outcomes. For example, the sentence copying activity used in this study demanded further receptive and productive focus on form and meaning and forced the participants to focus on each target item at least two more times (increasing the number of repetitions), which likely increased the learning outcomes (Folse, 2006; Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014). In addition, the sentence copying activity increased time-on-task, a factor that previous research has shown to produce superior learning outcomes (Hulstijn & Trompetter, 1998, pp. 191–200; Webb, 2005, Waring, 1997). From a practical point of view, as learners often spend only the minimum time needed to complete a task (Webb, 2005), a strong argument can be made in favor of adding vocabulary learning tasks that increase total study time and number of repetitions in natural teaching contexts (Folse, 2006; Webb, 2005). However, given that the findings of the interview data suggest that many students do not see the benefit of adding sentences to the cards, teachers can also consider the sentence copying task from an affective point of view. When students do not see any benefit to adding sentences, it could be that they pay little attention to the task of copying the sentences, thus reducing the efficacy of the activity.

In answer to the third research question, the qualitive data indicate that students view making word cards as a simple and quick process, and that they have a positive view of making and using word cards overall. First, students found the process of making word cards easy, and there were no problems during the treatment stage; students were not confused and were able to complete the treatment as instructed. Therefore, teachers can view word cards as a simple and easy to implement deliberate vocabulary learning technique that can be used in most language courses. Second and most important, students generally had a positive view of word cards for vocabulary building, supporting findings by Altiner (2011), Kuo and Ho (2012), Mondria and Mondria-de Vries (1994), Nikoopour and Kazemi (2014). The learners interviewed as part of this study viewed the process of making the cards as simple and efficient, but also stated that it required enough time, effort, and attention to detail to commit the new information to memory. In addition, the students favored the two-sided nature of the cards, something which adds affective support for the pedagogical perceived efficacy of word card design (Mondria & Mondria-de Vries, 1994; Nakata, 2008; Nation, 2013, pp. 438–451; Schmitt, 2008). However, none of the interviewees said that they enjoyed making word cards, so it can be assumed that the process is not intrinsically motivating; thus, the learning outcomes could be reduced (Deci & Ryan, 1985, pp. 16, 185, 225; Dörnyei, 2001, p. 28; Ng & Ng, 2015). In addition, some researchers compared learners' perceptions of paper word cards with digital cards or software and stated that learners preferred the latter (Emrah & Omur, 2010; Hung, 2015; Nakata, 2008). Given the mixed views noted above, and the fact that individual learner differences and vocabulary learning preferences exist in all teaching contexts (Barcroft, 2009; Kojic-Sabo & Lightbown, 1999, Mizumoto, 2010; Moir & Nation, 2002), it would be effective for teachers to consult with their students via pre-course surveys or action research regarding what type of word cards should be used.

CHAPTER 4

EXPERIMENT 2: STUDENT-MADE WORD CARDS VERSUS PRE-MADE WORD CARDS

Experiment 2 is also focused on the use of word cards as a deliberate vocabulary learning technique but differs from Experiment 1 in that three word card methods are compared. In addition, both a cross-sectional and semi-longitudinal design are utilized in Experiment 2.

Word cards usually take two forms: the traditional paper word card format, or the newer digital versions available on various websites and smartphone applications. In addition, as mentioned above, the amount and type of information that can be entered onto a card can vary considerably depending on learner or teacher preferences or goals. However, regardless of the forms word cards can take, a major similarity is that learners are usually expected to find and input the information on their cards themselves, something that is also true for vocabulary notebooks. A notable exception to this procedure comes with the increasing availability of pre-made digital word cards available online and through smartphone applications (Browne, 2014; Nakata, 2011). However, the study reported in this chapter is focused on traditional paper word cards, and the effect of students making their word cards versus studying from pre-made versions is examined.

Gaps in the Literature

As discussed in Chapter 2 and 3, while there is much support for the use of word cards, whether or not it is important for learners to make the cards or not is a severely overlooked issue. Only one study was found that compared the effect of making and studying word cards with studying from premade equivalents (Dodigovic, 2013) and that study focused on digital word cards and had many design flaws. Therefore, it is still not known if having learners make their own word cards is more effective than studying from premade versions. Experiment 2 addresses this important gap in the deliberate vocabulary study literature.

Therefore, the experiment described in this chapter is designed to fill an important gap in the literature regarding the processes and practice underlying vocabulary study from paper word cards. Specifically, I address the question of whether or not it is necessary for learners to make their own cards. Word cards are an effective deliberate study method and are generally viewed more favourably than word lists or notebooks (Ballance & Cobb, 2020, p. 324-325; Kuo & Ho, 2012; Nakata, 2008), but most research and discussions on word cards have not been focused on whether it is better for students to make their own cards or study from premade ones. While various theories and research relating to depth of processing (Craik & Lockhart, 1972), involvement load (Keating, 2008; Laufer & Hulstijn 2001), and task complexity (Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014; Pichette et al., 2011) indirectly support the practice of making word cards over studying from premade cards, no researchers have carefully compared the two approaches. The results of

Experiment 1 provided evidence that significant learning takes place when learners create word cards, but whether or not the time-consuming practice of making the cards is more beneficial than simply studying from pre-made equivalents for the same amount of time has not been investigated (Nation & Webb, 2011, p.42). This study is therefore the first in which making and studying from paper-based word cards is compared with studying from pre-made paper-based word cards.

Through a carefully controlled cross-sectional study, I first compare the immediate effects of making word cards versus studying from premade ones for the same amount of time. Then the learning outcomes associated with both word card methods after studying for an eight-week period are compared.

In addition, by comparing the learning outcomes associated with three different word card methods, this study investigates the appropriacy of using the Involvement Load Hypothesis to assess the potential learning outcomes of different word card designs. While numerous researchers have examined the vocabulary learning outcomes of activities with different task-induced involvement loads, none were found that applied the Involvement Load Hypothesis to word cards. Because the Involvement Load Hypothesis (ILH) has largely been used to evaluate incidental vocabulary learning tasks involving reading, writing, and speaking, little, if anything, is known about the importance of designing deliberate study methods in such a way as to manipulate the level of task-induced involvement load (Laufer & Hulstijn, 2001). To address this gap, the two controlled word card treatments and one more natural word card treatment were evaluated in terms of their individual task-induced involvement loads.

In addition to the vocabulary acquisition data, qualitative and quantitative data collected through observations, informal interviews, questionnaires, and artifact checking are used to examine student practices and perceptions concerning each method. The use of qualitative data collection methods offers holistic and triangulated data about the numerous individual learner variables involved in this study, many of which might have affected the learning outcomes.

Research Questions

The answers to the research questions below will (a) allow for a comparison of the learning outcomes attributable to self-made and premade word cards, (b) provide an indication of the suitability of using the ILH to assess potential learning outcomes of deliberate vocabulary methods, and (c) shed light on learner perceptions regarding word card methodology.

1. How much initial vocabulary learning occurs from the two controlled word card treatments: premade card treatment and student-made no-choice treatment?
2. Are either of the two controlled word card methods significantly more effective for developing short-term vocabulary knowledge?
3. Which of the three word card methods is superior in terms of vocabulary retention after repeatedly studying target words over a number of weeks?
4. To what extent does the level of task-induced involvement load predict the learning outcomes investigated in research questions 2 and 3?
5. How are the three experimental word card methods viewed by the participants?

To answer research questions 1 and 2, two types of pre-, post-, and delayed posttests were administered, which were the passive and active recall tests described in Experiment 1. A cloze test was not included in Experiment 2 due to time constraints. Also, qualitative data gained from questionnaires, observations, and interviews were used to answer research question 3. To answer research question 4, the experimental treatment variables involved in each study mode were assessed using the Involvement Load Hypothesis. Finally, research question 5 was answered by examining qualitative data gained from observation, surveys, and interviews.

Methods

Research Context

This experiment was conducted at the same two universities described in Experiment 1, but it was carried out in the following academic year with a different group of participants.

Participants

The participants were drawn from five intact classes; three from the private women's university and two from the private co-ed university described in Experiment 1. The initial sample consisted of 95 participants, but 11 were removed due to missing one or more stages of the experiment. Therefore, the final participant sample in this experiment consisted of 84 learners (69 female and 15 male students) aged between 18 and 21 (M = 18.6). The participants were taken from three of the courses described in

Experiment 1—Writing 2A, BCE, and EAP— but the students had enrolled in these courses one academic year later. In order to avoid repetition in terms of describing these three courses, only the specific participant details are described below.

Writing skills 2A (University 1). The participants came from two Writing 2A classes with a total of 43 female Japanese learners; however, four students were removed due to being absent from one or more of the tests. This left 39 participants aged between 19 and 21 (M = 19.3). Based on observations and informal interview data, the students had mixed levels of motivation for studying English, with most of them having no intention to study abroad for an extended period of time or having no clear reason for studying English. Their TOEIC scores ranged from 390 to 620, and pre-course VLT scores indicated that they had approximately 80% knowledge of the first 2,000 high-frequency words, 50% of the third 1,000, and less than 30% knowledge of the words in the Academic Word List.

Basic communicative English (University 1). The BCE course was made up of 19 female Japanese students, however, two students were removed for being absent from one or more of the tests, leaving 17 participants. All participants were aged between 18 and 19 (M = 18.06). Based on observations and informal interview data, the students had fairly high levels of motivation for studying English, with most stating that improving their English ability was one of their main goals during their university education. However, most of the students reported having no intention to study abroad for an extended period of time, and they had no clear reason for studying English. Their TOEIC scores ranged from 260 to 380, and based on the pre-course VLT administration, the participants had less than 80% knowledge of the first 2,000 high-frequency words, less than 40% of the third 1,000, and less than 20% knowledge of the words in the Academic Word List.

EAP 1 (University 2). This sample was taken from two intact classes with 33 students, however, due to missing one or more steps in the experiment, five students were removed from the study, leaving 28 participants; 15 male and 13 female learners aged 18. The participants were aged between 18 and 19 years old (M = 18.15). Based on observations and informal interview data, these students had fairly high levels of motivation for studying English, with approximately 50% of them stating that improving their English ability was one of their main goals during their university education. In addition, most of the students reported having an intention to study abroad for an extended period of time in the near future. However, as this was a compulsory class, the students were taking the course to reach the required number of English credits. Pre-course VLT scores indicated that the participants had approximately 70% of the first 2,000 high-frequency words, but less than 50% of the third 1,000 and AWL words .

In summary, five intact groups from two universities took part in this study. The total participant sample was 86 with a mean age of 18.9, but over 80% of the participants were female. The participants represented a convenience sample and were chosen predominantly due to ease of access.

Research Design

Experiment 2 is a mixed-methods study that uses a pretest-posttest design. The research design is illustrated in Figure 2.

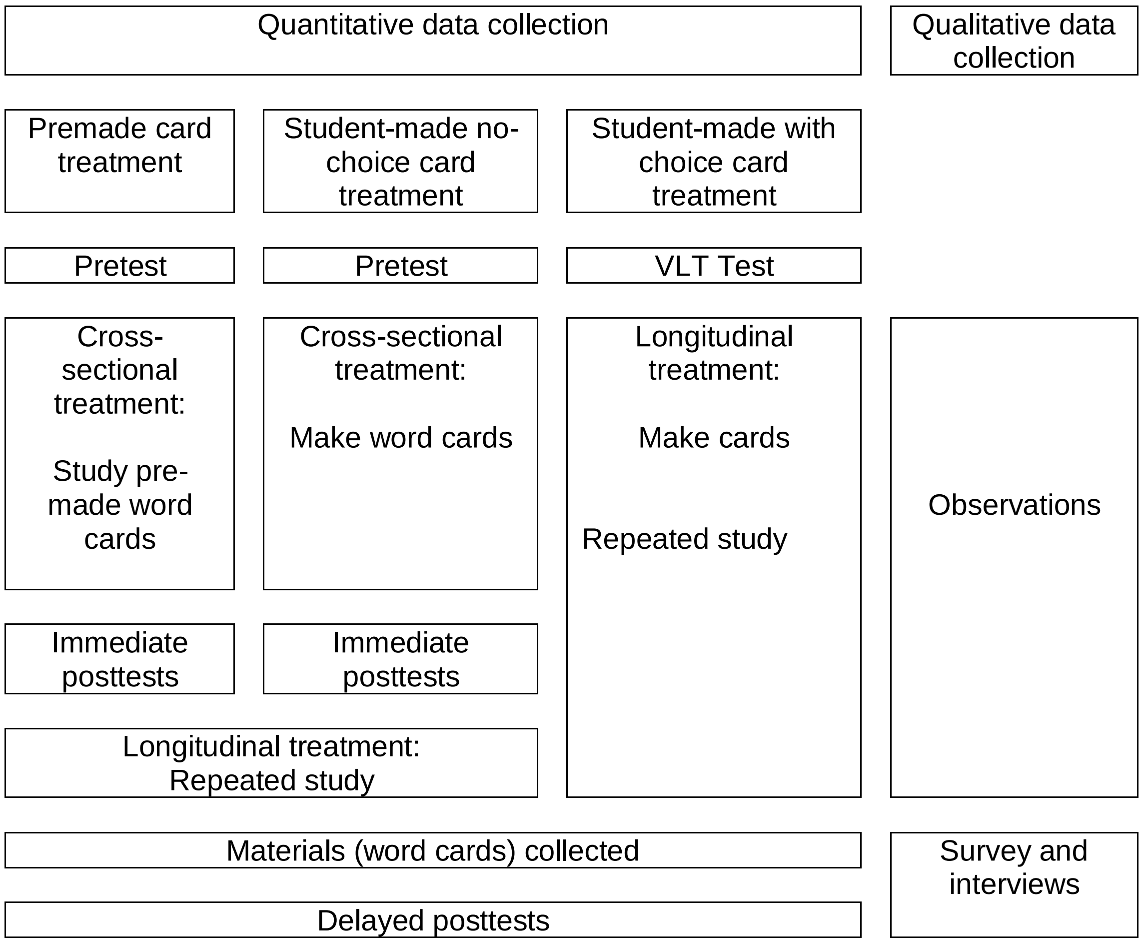

Figure 2. Research design for Experiment 2.

As with Experiment 1, due to wishing to avoid the use of a control group, and a desire to increase the experimental treatments' sample size, a counter-balanced design was adopted, and comparison items were used instead of adding a comparison group.

However, in Experiment 2, the participants studied words under three conditions, but due to time constraints, only two test types were used to assess the learning outcomes; active recall and passive recall. In addition, qualitative data from surveys, interviews, and observations were collected to gain more insight into individual learner factors that might have affected the results. Therefore, Experiment 2 utilizes a mixed-methods approach to fully answer the research questions and gain as much insight as possible.

Due to the research design, the three treatments not only differed in terms of their task-induced involvement loads, but also in the control of confounding variables such as the part of speech of the target words and sentence length. Therefore, in the following paragraphs, I first review the variables that have been controlled in each treatment, and then discuss how the treatments differ in terms of task-induced involvement load.

First, for all treatments, a number of confounding variables were controlled in order to ensure that the treatment variables were responsible for learning outcome differences found between the treatments. For example, in all three treatments, the physical word card format is the same, the type of information entered onto the cards is the same, and the words were taken from similar word frequency levels. In addition, as with Experiment 1, in both controlled word card treatments, the L2 target words were balanced in terms of parts of speech and word frequency level, their L1 translations were checked by multiple Japanese native speakers, and the example sentences were of similar length. Therefore, for the first two treatments, the premade card treatment and the student-made no-choice treatment, the main differences were (a) the way the target words and their L2 meanings were presented; on word cards or in a paper learner dictionary, and

(b) the style of learning; repeated receptive study involving only reading, or a combination of productive and receptive study involving reading, writing, and re-reading.

However, for the third treatment, because the participants self-selected both the L2 and L1 word forms and created their own sentences, there was more autonomy but less control of confounding variables such as the part of speech of the target words and sentence length. In addition, there are a number of other ways that the research design has resulted in other differences between treatments.

First, the number of encounters with target words, repetition, is a major factor that distinguishes the treatments. Under the pre-made card treatment, it was expected that the learners would carry out more repetitions with each target word compared to the student-made no-choice treatment and student-made with choice treatment, especially in the early stages because under the pre-made card condition, no time is needed to make the cards, so the participants can complete numerous repetitions, encountering the L2 word, their L1 meaning, and an example sentence several times in the time it takes to make the word cards in the student-made treatments.

In addition, while time-on-task was tightly controlled between the pretest and immediate posttest, this was not possible during the longitudinal part of the study. Therefore, the number of times the learners studied each word is unknown for the longitudinal part of the study, but because this variable exists in natural learning conditions, it is not a major concern. In addition, observation, interview, and questionnaire data were collected regarding this matter in order to better understand the effect of learner practices.

Second, due to the experimental design, each treatment had different scores on one or more of the three task-induced involvement load components of need, search, and evaluation. First, with regards to the *need* component, the task in all treatments was externally imposed by me, and the learners had to study and understand the words if they wished to earn a good course grade; therefore, there was a moderate need component of 1 (externally imposed need). However, the design of the treatments led to variability in terms of the second component, *search*. For the pre-made treatment, the L2 target words, their L1 meanings, and example sentences were provided; therefore, the search index was 0. However, under the student-made no-choice treatment, the learners engaged in a limited search by finding the words in the dictionary, reading the information, and transferring the required information to their word cards. This resulted in a search index of 1 for the student-made no-choice treatment. Finally, under the student-made with choice treatment, the students searched for unknown L2 target words from the word list given in class, and then they searched for and chose their L1 meaning using their own dictionaries before writing the chosen information onto their word cards. This resulted in a search component score of 2. Thus, each treatment's search component score ranged from 0 to 2. Finally, in terms of the *evaluation* component, while there was no difference between the premade treatment and the student-made no-choice treatment, both of which had an evaluation index of 0, the student-made with choice treatment had an evaluation index of 2 because the learners chose the best L1 meaning for the L2 target word and therefore evaluated the meaning of the L2 target word at the same time. When choosing the best L1 meaning, they often had many L1 choices and needed to evaluate which one most closely matched the target word. In addition, the learners looked up example sentences in their dictionaries or on the Internet, and then wrote a sentence. These two processes resulted in an evaluation score of 2. Table 13 provides a summary of the task-induced involvement loads for the three treatments.

Table 13. *Task-Induced Involvement Load Index*

Treatment	Need	Search	Evaluation	Total
Pre-made word cards with no choice of word or related information selection	1	0	0	1
Student-made word cards with no choice of word or related information selection	1	1	0	2
Student-made word cards with free choice of word and information selection information	1	2	2	5

In summary, in order to directly compare the premade treatment with the student-made no-choice treatment, a number of word-related confounding variables were controlled: The L1 target items, their L2 translations, and example sentences were exactly the same for both treatments. As a result, the participants have no choice concerning which L2 and L1 information to choose or write, and while this increased internal validity, it resulted in a somewhat unnatural word card treatment with lower ecological validity (Wegener & Blankenship, 2007, p. 167). This issue occurred because in a natural learning environment, after encountering the L2 form of an unknown word, learners usually search for their L1 meaning in a dictionary, and when multiple options are available, choose the one they think is best. In addition, when writing a sentence, learners either copy what they think is the best sentence from the choices provided in a dictionary or create an original sentence. However, none of these processes were available under premade treatment or student-made no-choice treatment due to an effort to reduce confounding variables and increase internal validity.

Therefore, the student-made with choice treatment, which represents a natural word card method, was added to this study. This treatment was deemed an ecologically valid treatment because it involved the types of processes learners normally go through when creating word cards. Under this condition, the learners created word cards after self-selecting the L2 target item by searching for and choosing what they thought was the best L1 meaning, and writing original sentences, all of which resulted in greater task-induced involvement. While this third treatment was more ecologically valid, it provided less control of confounding variables such as sentence length and parts of speech, so it was experimentally weaker.

Materials

Vocabulary items. The 40 L1 and L2 vocabulary items used in Experiment 2 were taken from the same pool of 71 items chosen and vetted for use in Experiment 1; 15 of these words were used in the first treatment condition, 15 for the second treatment, and 10 acted as comparison items. In addition, the participants self-selected 20 words per week for eight weeks from the 3,000-4,000-word frequency level of the COCA word list, the same list and frequency level that the other 40 items were taken from. In total, students created 160 word cards in this manner during the course of this study.

The 40 words selected by the researcher offered a sample that as closely as possible represented the range of vocabulary found in authentic texts in terms of part of speech; approximately 50% nouns, 22% verbs, 16% adjectives, and 8% adverbs (Johansson & Hofland, 1989). To match this ratio as closely as possible, each set of 15 target items contained 7 nouns (46.6%), 4 verbs (26.6%), 3 adjectives (20%), and 1 adverb (6.6%), and the comparison items included 5 nouns (50%), 3 verbs (20%), 1 adjective (10%) and 1 adverb (10%). However, the part of speech ratio of the students' self-selected words could not be controlled without altering the independent variable being examined; self-selection and its relationship to the involvement load hypothesis.

Word cards: Format and content. Experiment 2 involved the use of self-made word cards and pre-made word cards. The format and type of information of both the learner-made and pre-made cards in Experiment 2 was the same as those in Experiment 1. The two differences between the card types was the creation method—self-made or pre-made—and whether the participants had free choice in choosing the information to enter on to their cards—the L1 form of the L2 target item and example sentence.

Pre-made word cards. The learners were given 15 pre-made word cards that included the same information as in the dictionary style handout: the L1 word, the L2 target word, part of speech, and an example English sentence (see Appendix G). The pre-made word cards were created by three assistants under my supervision. The cards followed the format shown in the following example. As shown in the example, the premade cards, like the self-made version, included an example sentence. The pros and cons of including example sentences for the self-made cards were discussed in Experiment 1, and many of the benefits hold true under the pre-made treatment. However, there are differences in terms of the advantages offered, but these differences are the result of the main independent variable under examination, word card creation method.

Side A	Side B
efficient (adjective)	効率的な（形容詞） The capital city has a highly ______ public transport system.

In the pre-made word card treatment, there are fewer advantages to providing example sentences on the cards compared with the self-made treatment. For example, the inclusion of a sentence does not offer the productive element because no writing is required. In addition, if students choose not to read the sentences, there are no additional repetitions. However, based on the qualitative data analysis in Experiment 1, this outcome is unlikely, so these benefits likely hold for the pre-made card treatment. If the students read the sentences when studying from the cards, they can obtain the same information about how the word is used as in the self-made treatment. In addition, because they do not have to spend time writing the sentences, they have opportunities for more repetitions. Furthermore, the inclusion of these sentences was necessary to control the experimental variables. Unity in card design was necessary because it was important that both types of treatment provided the learners with the same information; otherwise, the treatments would have varied in more ways than those under investigation, and the lack of sentences in the pre-made treatment would have become a confounding variable. Overall, providing sentences on the pre-made cards offered similar learning benefits as the self-made treatment and ensured that the treatments differed only in the independent variable under investigation, creation method.

Overall, the inclusion of example sentences in both treatments offered the chance for additional repetitions while providing information about target word *use*.

Student-made no-choice word cards. For the second treatment, the participants were required to make 15 words cards under controlled conditions. As with Experiment 1, the learners were provided with 15 target items in a dictionary style booklet and 15 blank cards that they used to make the word cards. The participants were given both the pro forma and an example as described in Experiment 1.

Student-made free choice word cards. In the third treatment, the learners bought 200 blank cards and were given items from the 3,000-4000-word frequency level of the COCA, with the 30 target items used in the first two treatments and the 10 comparison items removed. The participants were then instructed to self-select 20 unknown items and make word cards every week for eight weeks. For each word card, the participants were instructed to find an unknown L2 word from the word list, and then search for the L1 meaning and an example sentence using their dictionaries or the Internet. They then created the word cards following the pro forma and example word card.

Instruments: Pretests and Posttests

Due to time constraints and research design issues, only the active and passive recall style tests were used for the pretests, immediate posttests, and delayed posttests in Experiment 2. Because the students self-selected the example sentences in the student-made with choice treatment, each participant's sentences differed; therefore, it was not possible to design a cloze test for each student. In addition, the inclusion of a third test type was not possible due to time constraints and the risk of test fatigue. The active and passive recall tests contained the 30 target items studied under the premade card and student-made no-choice treatments, plus the 10 comparison items. The pretests, immediate posttests, and delayed posttests were made up of the same 40 items, but the word order was changed. In order to test the learners' knowledge of words that were self-selected at the delayed posttest stage, a student-generated test was used that tested 10 words through an active recall test and 10 items through a passive recall test. This spilt passive-active recall test design was chosen because it would have taken too much time to test all target words and comparison items both actively and passively and doing so would have increased the likelihood of test fatigue.

Procedures

The VLT was administered in order to gauge the participants' vocabulary level in order to ensure that the target vocabulary would be unknown and that an appropriate word list was chosen for the student-made with choice treatment. However, in order to establish a baseline knowledge of the target and comparison items, the students took the pretests in the first week of the semester.

Then two weeks later, in week three, the treatment stage began. In class, half of the participants (Group 1) were given 15 minutes to make 15 word cards, while Group 2 had 15 minutes to study the same words using pre-made cards. The treatments were then reversed; the Group 1 participants studied a further 15 words using pre-made word cards, while the Group 2 participants made 15 word cards containing the same items and information. This treatment order was used to partially control for word effects and order effects. Then, directly after the end of the treatment phase, the students took an immediate posttest.

Next, as part of the longitudinal design, for eight weeks (after class in week three to the end of week 11), students created the word cards used in the third treatment. To do this, participants self-selected 20 words each week from the 3,000-4,000 word frequency level of the COCA and created word cards using their dictionaries. There was no pretest, but it is unlikely that the learners knew these words based on (a) the participants' VLT scores, (b) the scores on the above-mentioned pretests, and (c) the fact that clear instructions were given to select unknown words. The self-selected and self-made cards in the third treatment included the same type of vocabulary information included in the other two treatments: the target L2 word, the L1 meaning, and an example sentence.

Then for eight weeks, beginning of week four to the end of week 11, the students put all three types of word cards together and studied them under normal conditions. They were instructed to study all of their cards regularly outside of class. In addition, approximately ten minutes of class time each week was devoted to word card study, either individually or in pair or group-based activities. After the eight-week period, I collected and retained all of the word cards. Two weeks later, the participants took the delayed posttests. The 30 target items studied under the two teacher-selected word treatments and the comparison items were tested using the same tests used for the pretest and immediate posttest, but the order of the items was changed. The 20 words self-selected by the students were tested using a peer-generated written test that tested half the words actively and half passively. The students were paired with a partner, they gave their cards to their partner, and the partner randomly selected 20 words and wrote 10 of the words in Japanese (active recall test) and 10 items in English (passive recall test). It was not possible to test all 20 items both passively and actively because it would have taken too much class time and increased the risk of test fatigue.

Data Analysis

The quantitative data for Experiment 2 were analyzed using a one-way repeated-measures MANOVA, a one-way repeated-measures ANOVA, and a two-way repeated-measures ANOVA. All analyses were carried out using IBM SPSS Statistics 21.

Like Experiment 1, a repeated-measures analysis was chosen because all participants experienced the same treatments, and the pretest-posttest design meant that all participants and were tested over three time points. The counterbalanced design accounted for and corrected individual participant differences that might have impacted the outcomes of the experiment (Laerd Statistics, 2018; Verma, 2016).

In addition to the above, qualitative data from observations, questionnaires, and interviews were collected and analyzed to offer further insights into perceptual and procedural issues that might have affected the results. Interviews and questionnaires were used to investigate differences in individual learners' study time and strategies, examine the participants' perceptions of each method, and investigate whether one method was viewed more favorably than another. The questionnaire included a combination of Likert scale questions and open-ended questions that the participants could answer in their L1, Japanese. The questions were focused on how the learners felt about each method in terms of its efficiency, efficacy, interest level, and how often they studied from the cards made in each condition. The interviews were conducted to garner more detailed information about the same issues. The qualitative data provided insights about reasons for the learning outcomes.

Results: Learning From Premade and Self-Made Word Cards

Analysis of Comparison Items

The first analysis, a one-way repeated-measures MANOVA using SPSS Statistics on the pretest means, was carried out to investigate if there were any statistically significant differences between the three groups on the two dependent measures at the start of the study. The factor was treatment (three levels: pre-made word card treatment, self-made word card treatment, and control), and the two dependent variables were the passive recall and active recall tests. As described below, before analyzing the data, the assumptions of the analysis were checked to assess if MANOVA analysis was suitable.

First, the passive and active recall tests were measured on an interval scale. Second, all participants were measured at the same time point and received all treatments. In addition, the participants were not able to influence the scores of other participants because there was no interaction between students during the treatment and test stages. Next, the final sample of 83 participants was of a sufficient size and complete data for all of the participants was ensured by removing the participants who were not present for any of the treatment or testing stages. In addition, the linear relationship between each pair of dependent variables for all combinations of groups on the two independent variables was confirmed after plotting a scatterplot matrix. A weak linear relationship was confirmed between the dependent pretests and the two posttests, and there was a strong linear relationship between the immediate and delayed posttests measures.

Next, the presence of multivariate outliers was checked by calculating and analyzing the Mahalanobis Distances (Field, 2018, pp. 422–423), and comparing them to a chi-square distribution with the same degrees of freedom to compute the probability variable (Laerd Statistics, 2019). The Mahalanobis Distances analysis identified 15 possible outliers at the pretest stage; cases that were above the critical value of the chi-square distribution of 15 (Field, 2018, p. 422), but only four whose probability variable values were less than .001. In addition, the presence of univariate outliers was examined using Stem-and-Leaf Plots. In total, the analysis identified 15 univariate outliers. However, after considering the research design and looking at the descriptive statistics, it was felt that removing these outliers was undesirable because they represented natural variations in word knowledge before the treatment began and most of the values were not extreme. However, as a precaution, non-parametric Friedman tests were carried out, the results of which supported those of the MANOVA analysis in terms of lack of statistical significance for the treatment.

Next, the normality of each dependent variable for each of the independent variable groups was tested and unfortunately the Shapiro-Wilk test was significant for all dependent measures and z-skewness and z-kurtosis was > 3.29 for many of the dependent measures, indicating that the assumption of normality was violated. However, the Q-Q plots showed that normality did not stray far from the expected values, and as mentioned, Blanca et al. (2017) concluded that non-normality is not an important issue where ANOVA and MANOVA analyses are concerned. Taking these views into consideration along with the observations of the Q-Q plots, it was decided that the violation was not too severe. Finally, the correlation between the dependent variables was checked using Pearson correlations, the results of which showed that all but one pair ($r = .26$) of the pretest scores were correlated with each other at between $r = .31$ and $r = .87$, indicating that there was sufficient correlation but no multicollinearity.

Before discussing the results of the MANOVA analysis, the descriptive statistics for the pretest data are shown in Table 14.

Table 14. Descriptive Statistics for the Three Treatments on the Active Recall Pretest and Passive Recall Pretest

	Premade card treatment	Self-made card treatment	Control
Active recall pretest			
M	0.24	0.23	0.24
SD	0.77	0.67	0.64
SE	0.08	0.07	0.07
95% CI	[.07, .41]	[.08, .37]	[.10, .38]
Skewness	3.75	2.96	2.37
SES	0.26	0.26	0.26
Kurtosis	14.33	7.95	3.86
SEK	0.52	0.52	0.52
Passive recall pretest			
M	0.41	0.43	0.36
SD	0.98	1.04	0.77
SE	0.11	0.11	0.08
95% CI	[.20, .62]	[.20, .66]	[.19, .53]
Skewness	2.39	2.39	1.69
SES	0.26	0.26	0.26
Kurtosis	5.15	4.90	.877
SEK	0.52	0.52	0.52

Note. *N* = 83. CI = Confidence interval. 15 target items (maximum 30 points) and 10 comparison items (maximum 20 points).

The repeated-measures MANOVA indicated that there was no significant main effect for treatment at the pretest stage, Wilks $\Lambda = .99$, $F(4, 79) = .19$, $p = .94$, partial $\eta^2 = .02$, observed power = .09. Thus, there was no significant differences between the three groups on the dependent measures before the treatments began.

Next, in order to investigate if there were significant gains in the mean scores on the comparison items over time, two one-way repeated-measures ANOVAs were carried out. The repeated-measures ANOVA analysis was chosen over a repeated-measures MANOVA because the Pearson correlation analysis showed that many of the active and passive recall measures were not appropriately correlated. Four test pairs were correlated below .30 ($r = .16$, $r = .21$, $r = .23$, and $r = .29$), and two were correlated above .70 (r =

.78. and $r = .79$). Thus, an assumption of repeated-measures MANOVA analysis was violated, so one-way repeated-measures ANOVAs were carried out.

Before performing the analysis, the assumptions of the repeated-measures ANOVA were confirmed. First, in terms of data types, the dependent variable was measured at the interval level, the independent variable consisted of three related groups; all participants were measured at all time points and received the same treatments, and there was a suitable sample size. Next, the presence of outliers was checked using Boxplots and 15 outliers were identified. Removing these outliers was not ideal as they represented real participant vocabulary knowledge measures, so to check if it was possible to keep these outliers, nonparametric Friedman tests were performed and compared to the ANOVA outputs, and the results indicated that the outliers did not affect the ANOVA results. Therefore, it was decided that the analysis could be carried out with data from all 83 participants. Next, the Shapiro-Wilk test was significant for all dependent measures and z-skewness and z-kurtosis was > 3.29 for many of the active and passive recall dependent measures. However, as discussed earlier, Blanca et al. (2017) and Khan and Raynor (2003) concluded that non-normality is not an important issue where ANOVAs are concerned. Finally, because Mauchly's test of sphericity was significant, the Greenhouse-Geisser correction was used.

The factor for each ANOVA was time of testing (three times: pretest, immediate posttest, and delayed posttest) and the dependent variable was either the active or passive recall measure of comparison item vocabulary knowledge. For the first repeated-measures ANOVA, the dependent measure was the active recall test, and for the second it was the passive recall test. A repeated-measures ANOVA indicated that there was no main effect for time for the active recall measure of the comparison items, $F(2, 164) = 1.61$, $p = .20$, partial $\eta^2 = .02$, observed power = .34. Second, a repeated-measures ANOVA with a Greenhouse-Geisser correction indicated no main effect for time for the passive recall test mean scores of the comparison items, $F(1.59, 130.09) = .19$, $p = .78$, partial $\eta^2 = .002$, observed power = .08.

The results of the one-way repeated-measures ANOVAs showed no significant difference between the mean scores on the comparison item measures between the pretest, posttest, and delayed posttest stages, meaning that no learning took place. This was a positive result because it indicated that the participants did not learn from the test and were not exposed to the words outside of the experiment. Therefore, this eliminated the two major possible confounding variables that could not be controlled in the design; as a result, it can be safely assumed that any learning that takes place under the two word card treatment conditions can be attributed to the treatment.

Analysis of Treatment Items

In order to investigate if there was (a) statistically significant vocabulary gains as a result of making word cards, and (b) significant attrition after not studying the words for two weeks, two-way repeated-measures ANOVA analyses were carried out; one for each dependent measure of vocabulary knowledge. A repeated-measures ANOVA analysis was chosen over a repeated-measures MANOVA because the Pearson correlation analysis showed that many of the active and passive recall measures were not appropriately correlated. Six test pairs were correlated below .30 (r = .06, r = .17, r = .18, and r = .19, r = .27, r = .28), and four were correlated at above .80.

The factors for the repeated-measures ANOVAs were treatment (two levels: premade word cards and self-made) and time of testing (three times: pretest, immediate posttest, and delayed posttest); the dependent variable was one of the two measures of vocabulary knowledge; active recall and passive recall.

The assumptions underlying the two-way repeated-measures ANOVA were checked. The dependent variable was measured at the interval level, and the two within-subject factors (time and treatment) consisted of two categorical related groups; the same participants are present in both groups (Laerd Statistics, 2018). In addition, an examination of the box plots identified ten outliers on the active pretest measure and 14 on the passive pretest but there were no outliers on any of the posttests or delayed posttests. Therefore, given the previous pretest analysis and lack of outliers on posttests, it appeared that this violation was not so severe. The normality of the dependent variable in each combination of the related groups was problematic; the Shapiro-Wilk test was significant for many combinations, but the Q-Q plots showed that the normality did not stray far from the expected values. Further analysis of the data indicated that-skewness was > 3.29 and z-kurtosis > 3.29 for the pretest active and passive recall dependent measures, but < 3.29 for all of the posttests. Given that most distributions were normal, and as Blanca et al. (2017) and concluded that non-normality is not an important issue where ANOVAs is concerned, the violations were not considered problematic and the outliers were retained. However, due to some violations, as a precautionary measure, further analysis was carried out after each ANOVA using the non-parametric Friedman test, the results of which mirrored those of the ANOVAs in terms of the statistical significance of differences in the dependent measure means, so only the ANOVA results are reported. Finally, when Mauchly's test of sphericity was significant, the Greenhouse-Geisser correction was used.

Active recall test results. A two-way repeated-measures ANOVA was used with the within factors being time (three levels: pretest, immediate posttest, and delayed posttest) and treatment (two levels: premade and self-made no-choice) and the dependent measure was the active recall test. The descriptive statistics presented in Table 15 are discussed below.

As seen in Table 15, while there was little difference between the two treatments' active recall means at pretest, the premade word card treatment mean was much higher on the immediate posttest and slightly higher on the delayed posttest. The difference between the treatments was especially noticeable at the posttest stage given that the premade mean was 18.06, while the self-made mean was 12.38. Given the partial scoring method explained above, the mean difference of almost 6 means a difference of approximately 3 words out of a total of 15 (20% extra word knowledge). However, while the premade card treatment active recall word knowledge means decreased from posttest to delayed posttest, those of the self-made card treatment showed small gains. This result is surprising because it shows an unexpected difference between the groups. Although the premade treatment mean is higher than the self-made treatment at both posttest stages, the premade treatment knowledge decreased between posttest and delayed posttest—after eight weeks of studying the cards before having a two-week period of no study—while that of the self-made group increased.

Table 15. *Descriptive Statistics for the Active Recall Test for Each Treatment at Each Time Point*

	Premade card treatment	Self-made card treatment
Pretest		
M	0.24	0.23
SE	0.08	0.07
95% CI	[.07, .41]	[.08, .37]
SD	0.77	0.67
Skewness	3.75	2.96
SES	0.26	0.26
Kurtosis	14.33	7.95
SEK	0.52	0.52
Posttest		
M	18.06	12.38
SE	0.89	0.99
95% CI	[16.28, 19.84]	[10.42, 14.35]
SD	8.14	9.01
Skewness	-0.20	0.59
SES	0.26	0.26
Kurtosis	-1.23	0.81
SEK	0.52	0.52
Delayed posttest		
M	14.90	13.79
SE	1.02	1.10
95% CI	[12.88, 16.93]	[11.60, 15.99]
SD	9.26	10.03
Skewness	0.21	0.23
SES	0.26	0.26
Kurtosis	-1.28	-1.41
SEK	0.52	0.52

Note. *N* = 83. CI = Confidence interval. 15 target items: maximum score is 30.

Next, as determined by a two-way repeated-measures ANOVA, there were statistically significant main effects for time, treatment, and time-treatment in terms of active recall knowledge. First, the repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that the mean active recall scores differed significantly between time points, $F(1.61, 132.32) = 206.39$, $p < .001$, partial $\eta^2 = .72$, observed power = 1.00. There was also a significant main effect for treatment, $F(1, 82) = 35.84$, $p < .001$, partial $\eta^2 = .30$, observed power = 1.00. Finally, the repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that the time-treatment interaction was significant, $F(1.74, 142.81) = 26.22$, $p < .001$, partial $\eta^2 = .24$, observed power = 1.00. The results indicated that the mean scores on the active vocabulary knowledge measure changed over time, and that they differed between the two treatments for at least one of the three time points.

Given the statistically significant effect for treatment and treatment-time, further analyses were conducted to investigate whether (a) the mean changes between time points were statistically significant for each treatment, and (b) the mean differences between the treatments were significant at posttest and delayed posttest. Two separate one-way repeated-measures ANOVA analysis were carried out, the factor for each being time (three levels: pretest, posttest, and delayed posttest), and the dependent measure was the active recall test.

The first repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that mean active recall scores differed significantly between time points for the premade word card treatment, $F(1.79, 146.41) = 236.68$, $p < .001$, partial $\eta^2 = .74$, observed power = 1.00. The second repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that there was also a significant difference over time for the self-made treatment, $F(1.84, 151.02) = 116.94$, $p < .001$, partial $\eta^2 = .59$, observed power = 1.00.

In order to investigate whether each of the changes between time points were statistically significant for each treatment, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at < .017). The results of the pairwise comparisons are presented in Table 16.

Table 16. *Pairwise Comparisons for the Active Recall Test for Each Treatment at Each Time Point*

	Premade card treatment	Self-made card treatment
Posttest to pretest		
Mean difference	17.82	12.16
SE	0.87	0.98
95% CI	[15.69, 19.95]	[9.762, 14.55]
p-value	< .001	< .001
Delayed posttest to pretest		
Mean difference	14.66	13.57
SE	0.99	1.09
95% CI	[12.22, 17.10]	[10.90, 16.2]
p-value	< .001	< .001
Delayed posttest to posttest		
Mean difference	-3.16	1.41
SE	0.73	0.84
95% CI	[-4.95, -1.37]	[-.64, 3.46]
p-value	< .001	.29

Note. N = 83; CI = Confidence interval. 15 target items: maximum difference is 30.

As seen in Table 16, the increased mean from pretest to both posttests was statistically significant for the premade group, but so was the attrition from the posttest to the delayed posttest. The self-made treatment also displayed significant gains from pretest to posttest and delayed posttest; the gains from the posttest to the delayed posttest were not significant. Thus, both word card treatments were successful in terms of building active recall vocabulary knowledge. In addition, there were positive and negative results for each treatment; the premade treatment was more effective overall but suffered greater attrition, while the self-made treatment resulted in no significant attrition.

Next, in order to investigate whether the mean scores of each treatment differed significantly at each time point, three one-way repeated-measures ANOVAs were carried out with the factor being treatment (two levels; premade, self-made no-choice) and the dependent measure being the active recall tests. The results indicated that the mean active recall scores of the two treatments did not differ significantly at the pretest stage, Wilks $\Lambda = 1.00$, $F(1, 82) = .04$, $p = .85$, partial $\eta^2 = .00$, observed power = .05, or at the delayed posttest, Wilks $\Lambda = .97$, $F(1, 82) = 2.35$, $p = .13$, partial $\eta^2 = .28$, observed power = .33. However, the treatment differed significantly at the immediate posttest stage, Wilks $\Lambda = .60$, $F(1, 82) = 54.60$, $p < .001$, partial $\eta^2 = .40$, observed power = 1.00. The results of the pairwise comparisons performed with a Bonferroni correction for multiple comparisons are presented in Table 17 (p is significant at < .017).

Table 17. *Pairwise Comparisons between the Treatments for the Active Recall Test at the Three Time Points*

	Premade card treatment to self-made card treatment
Pretest	
Mean difference	0.01
SE	0.06
95% CI	[-.11, .14]
p	.85
Posttest	
Mean difference	5.67
SE	0.77
95% CI	[4.15, 7.20]
p	< .001
Delayed posttest	
Mean difference	1.11
SE	0.72
95% CI	[-.33, 2.55]
p	= .13

Note. *N* = 83. CI = Confidence interval. 15 target items: maximum difference is 30.

Finally, in order to compare the active recall learning outcomes of the two experimental treatments with the third word card treatment, the most natural method, looking at the descriptive statistics for each is useful. As seen in Table 18, in percentage terms, the third treatment, the natural word card treatment, resulted in higher learning outcomes than the other two treatments at the delayed posttest stage.

Table 18. *Descriptive Statistics for the Active Recall Test Mean Percentages for Three Treatments on the Delayed Posttest*

	Premade card treatment (15 words)	Self-made no-choice card treatment (15 words)	Self-made with choice card treatment (10 words)
M (% of maximum)	49.68%	45.98%	64.40%
SE	3.39	3.67	2.76
95% CI	[42.93, 56.42]	[38.68, 53.29]	[58.91, 69.89]
SD	30.88	33.44	25.15

Note. *N* = 83. CI = Confidence interval.

Because the words and card information studied under the third condition were self-chosen by the students, rigorous statistical analysis was not possible due to (a) an inability to administer a pretest of these words, and (b) the lack of control for confounding variables such as choice of L1 translation, length of example sentences, and balance of parts of speech. However, by running a one-way repeated-measures ANOVA on the delayed posttest scores (using percentages of correct answers) and by comparing the descriptive data and pairwise comparisons, the effect of the differing levels of task-induced involvement load could be gleaned. The percentage of correct answers was used because while the two experimental treatments tested 15 words both actively and passively, the natural treatment measured 10 words actively and 10 passively, making a direct comparison of the mean scores impossible.

A one-way repeated-measures ANOVA was carried out where the within-subjects factor was treatment (three levels: premade, self-made no-choice, self-made with choice), and the dependent measure was the active recall delayed posttest. The repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that active recall means differed significantly between the three treatments at the delayed posttest, $F(1.79, 147.06) = 24.16$, $p < .001$, partial $\eta^2 = .23$, observed power = 1.00.

Next pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at < .017), the results of which are presented in Table 19.

Table 19. Pairwise Comparisons for the Active Recall Test Mean Percentages for the Three Treatments on the Delayed Posttest

	Self-made with choice treatment to Premade Treatment	Self-made with choice treatment to Self-made no-choice treatment	Premade treatment to Self-made no-choice treatment
M difference (%)	14.72%	18.41%	3.69%
SE	2.72	3.22	2.41
95% CI	[8.08, 21.36]	[10.05, 26.28]	[-9.59, 2,20]
p-value	< .001	< .001	= .388

Note. *N* = 83. CI = Confidence interval.

The pairwise comparisons indicate that the differences between the self-made with choice and the two controlled experimental conditions (premade and self-made with no choice) differed significantly on the delayed posttest. As stated above, due to the lack of controls regarding the third treatment, and given the use of percentage data, these results must be interpreted with caution. In addition, the high gaps between the lower and upper bound confidence interval are a major concern as they indicate that the effect of each condition varied considerably. In addition, the high standard deviations show that the treatment was not effective for all students; some earned very low scores while others earned high scores. However, taken together, the means and the ANOVA results indicate that the natural treatment, which had the highest task-induced involvement load, resulted in superior active recall learning outcomes. However, there was no significant difference between the premade card treatment (lowest task-induced involvement load) and the self-made treatment (middle ranked task-induced involvement load). Analysis of the qualitative data provided further insights into why these long-term differences exist and triangulated the findings.

Passive recall test results. The first analysis involved running a two-way repeated-measures ANOVA with two within-subject factors; Treatment (two levels: Premade and self-made no choice) and Time (three levels: pretest, immediate posttest, and delayed posttest). The dependent measure was the passive recall test.

The descriptive statistics are presented in Table 20; the mean scores for both treatments were almost the same at the pretest stage, but the premade word card treatment means were higher at both posttest stages. However, the difference between the means of both groups was quite small at both posttest stages. In addition, there was attrition from the posttest to the delayed posttest for both treatments. The results showed considerable learning from the pretest to the posttest and delayed posttest, but a small amount of attrition occurred between the posttest and delayed posttest; thus, even after eight weeks of repeatedly studying these words under both treatment conditions, when no study is carried out for two weeks, passive vocabulary knowledge attrition occurs. In terms of the self-made treatment, these results are surprising because the participants made slight gains in active recall knowledge from the posttest to the delayed posttest.

Table 20. Descriptive Statistics for the Passive Recall Test for Each Treatment at Each Time Point

	Premade card treatment	Self-made card treatment
Pretest		
M	0.41	0.43
SD	0.97	1.04
SE	0.11	0.11
95% CI	[.20, .62]	[.21, .66]
Skewness	2.40	2.39
SES	0.26	0.26
Kurtosis	5.15	4.90
SEK	0.52	0.52
Posttest		
M	19.45	16.99
SD	7.31	8.24
SE	0.80	0.90
95% CI	[17.85, 21.04]	[15.19, 18.79]
Skewness	-0.29	0.00
SES	0.26	0.26
Kurtosis	-1.07	-1.23
SEK	0.52	0.52
Delayed posttest		
M	18.73	16.02
SD	7.87	8.92
SE	0.86	0.98
95% CI	[17.02, 20.45]	[14.08, 17.97]
Skewness	-0.09	0.03
SES	0.26	.26
Kurtosis	-1.12	-1.20
SEK	0.52	0.52

Note. *N* = 83. CI = Confidence interval. 15 target items: maximum score is 30.

To investigate whether the differences discussed above were significant, a two-way repeated-measures ANOVA analysis with a Greenhouse-Geisser correction was performed. There was a significant main effect for time; $F(1.64, 134.50) = 365.09$, $p < .001$, partial $\eta^2 = .82$, observed power = 1.00, treatment; $F(1, 82) = 26.95$, $p < .001$, partial $\eta^2 = .25$, observed power = 1.00, and time-treatment; $F(2, 164) = 11.64$, $p < .001$, partial $\eta^2 = .12$, observed power = .99. The results indicated that the mean scores on the passive vocabulary knowledge measure changed over time and differed between the two treatments for at least one of the three time points.

In order to investigate where the significant differences lie, further analyses were needed. First, to investigate whether the mean changes between time points were statistically significant for each treatment, two separate one-way repeated-measures ANOVAs were carried out with the factor being time (three levels: pretest, immediate posttest, and delayed posttest) and the dependent measure being passive recall. The repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that mean passive recall scores differed significantly between time points for the premade word card treatment, $F(1.73, 142.19) = 414.65$, $p < .001$, partial $\eta^2 = .83$, observed power = 1.00, and for the self-made treatment, $F(1.76, 144.10) = 233.59$, $p < .001$, partial $\eta^2 = .74$, observed power = 1.00.

Given the above results, it was necessary to examine whether the changes between time points were statistically significant for each treatment. Therefore, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at $< .017$), the results of which are presented in Table 21.

Table 21. Pairwise Comparisons for the Passive Recall Test for Each Treatment at Each Time Point

	Premade card treatment	Self-made card treatment
Posttest to pretest		
Mean difference	19.04	16.55
SE	0.78	0.90
95% CI	[17.13, 20.95]	[14.35, 18.75]
p	< .001	< .001
Delayed posttest to pretest		
Mean difference	18.32	15.59
SE	0.85	0.97
95% CI	[16.25, 20.40]	[13.23. 17.95]
p	< .001	< .001
Delayed posttest to posttest		
Mean difference	-0.71	-0.96
SE	0.59	0.69
95% CI	[-2.159, .737]	[-2.65, .72]
p	.70	.50

Note. *N* = 83. CI = Confidence interval. 15 target items: maximum difference is 30.

As seen in Table 21, the increases in means from the pretest to the posttest and delayed posttest were statistically significant for both treatments, but the attrition from the posttest to the delayed posttest was not; thus, both word card treatments led to gains in passive recall vocabulary knowledge in the short-term and long-term. Slight statistically insignificant attrition occurred. While the premade treatment was the most effective, there was not a large gap in learning outcomes between the two treatments.

Next, in order to investigate whether the mean scores of each treatment differed significantly at each time point, three one-way repeated-measures ANOVAs were performed. The factor was treatment (two levels; premade, self-made no-choice) and the dependent measure was passive recall test. The repeated-measures ANOVA indicated that while the mean passive recall pretest scores of the two treatments did not differ significantly at the pretest; Wilks $\Lambda = 1.00$, $F(1, 82) = .07$, $p = .80$, partial $\eta^2 = .00$, observed power = .057, the two treatment means differed significantly on the posttest, Wilks Λ = .85, F(1, 82) = 14.65, p < .001, partial η^2 = .15, observed power = .97, and on the delayed posttest, Wilks Λ = .77, F(1, 82) = 24.60, p < .001, partial η^2 = .23, observed power = .99. Next, pairwise comparisons for each test time were performed with a Bonferroni correction for multiple comparisons (p is significant at < .025). The pairwise comparison data for each test time are shown in Table 22.

Table 22. *Pairwise Comparisons between the Treatments on the Passive Recall Test at Each Time Point*

	Premade card treatment to self-made card treatment
Pretest	
Mean difference	-0.02
SE	0.09
95% CI	[-.21, .16]
p	.79
Posttest	
Mean difference	2.46
SE	0.64
95% CI	[1.18, 3.73]
p	< .001
Delayed posttest	
Mean difference	2.71
SE	0.55
95% CI	[1.62, 3.80]
p	< .001

Note. *N* = 83. CI = Confidence interval. 15 target items: maximum difference is 30.

While there was no significant difference at the pretest stage, the premade treatment resulted in significantly higher passive recall learning outcomes at the post test and delayed posttest stages. The mean difference increased slightly between the two treatments from posttest to delayed posttest with the premade group having a significantly higher mean. Therefore, unlike the active recall measure, in the long-term, the premade treatment was superior to the self-made treatment in terms of developing passive recall knowledge of target items. This result was surprising because the premade treatment had a lower task induced involvement load than the self-made treatment.

Finally, further analysis was carried out to compare the learning outcomes of the two experimental treatments with the third word card treatment; the natural word card method. As seen in Table 23, in percentage terms, the self-made with choice card treatment resulted in higher learning outcomes than the other two treatments at the delayed posttest stage.

Table 23. *Descriptive Statistics for the Passive Recall Test Mean Percentages for the Three Treatments on the Delayed Posttest*

	Premade card treatment (15 words)	Self-made no-choice card treatment (15 words)	Self-made with choice card treatment (10 words)
M (% of maximum)	62.07%	52.93%	73.11%
SD	26.16	29.58	22.91
SE	2.89	3.27	2.53
95% CI	[56.32, 67.82]	[46.43, 59.42]	[68.07, 78.14]

Note. *N* = 83. CI = Confidence interval.

Next, in order to investigate if these differences were statistically significant, a one-way repeated-measures ANOVA was carried with the factor being treatment (premade, self-made no-choice, self-made with choice) and the dependent measure being the delayed posttest passive recall measure. The analysis was carried out using mean percentage of correct answers on the delayed posttest scores because while the two experimental treatments tested 15 words both actively and passively, the natural treatment measured 10 words actively and 10 passively. As a result of a lack of rigorous control in the word card making process, and given the use of percentages, the results of the statistical analysis need to be interpreted with caution.

As shown in Table 24, the repeated-measures ANOVA with a Greenhouse-Geisser correction indicated that mean passive recall scores differed significantly between the groups, $F(1.52, 123.06) = 26.83$, $p < .001$, $\eta^2 = .25$, observed power = 1.00; thus, the natural treatment and the two controlled experimental conditions differed significantly on the delayed posttest. Next, to investigate where the statistically significant differences were, pairwise comparisons for each test time were performed with a Bonferroni correction for multiple comparisons (p is significant at $< .017$). As seen in Table 24, the self-made with choice treatment resulted in significantly better passive recall knowledge than the other two treatments. In addition, as discussed earlier, the premade treatment produced better passive knowledge than the controlled self-made card treatment.

Table 24. *Pairwise Comparisons for the Passive Recall Test Mean Percentages for the Three Treatments on the Delayed Posttest*

	Self-made with choice treatment to Premade Treatment	Self-made with choice treatment to Self-made no-choice treatment	Premade treatment to Self-made no-choice treatment
M % difference	11.04%	20.18%	9.15%
SE	3.02	3.21	1.84
95% CI	[3.65, 18.52]	[12.32, 28.04]	[4.65, 13.65]
p	= .001	< .001	< .001

Note. *N* = 83. CI = Confidence interval.

However, due to (a) the lack of controls regarding the third treatment, (b) the use of percentage data, and (c) the large gap between the lower and upper bound confidence levels, these statistical results must be interpreted with caution. Nevertheless, given the higher percentage value for the natural treatment, and considering that the 10 words tested were randomly selected from a pack of 220 words, they indicated that the natural treatment, which had the highest task-induced involvement load resulted in superior learning outcomes. However, the means of the premade treatment were significantly higher than that of the self-made treatment, a surprising result because the premade treatment had a lower task induced involvement load.

Results of Qualitative Data Analysis

The results of the qualitative data analysis highlighted key affective issues regarding word cards in general, and more specifically about the types of word card methods used in this study. The qualitative data analysis indicated that the participants had a positive view of using word cards overall, but they preferred studying from premade word cards.

General perceptions of word cards. Most participants displayed very positive views toward using word cards for deliberate vocabulary study. For example, 66 out of the 74 respondents (89.2%) rated word cards as effective (M = 4.6 out of 5.0) when answering the 5-point Likert-scale question *How effective do you think word cards are for learning new vocabulary?* Specifically, 75.7% of the participants (56 students) rated word cards as *Extremely effective* (5), 13.5% (10 students) chose *Very effective* (4), and 10.8% (8 students) selected *Somewhat effective* (3), and no students chose the lowest two scale items; *Not very effective* (2), and *Not at all effective* (1).

When asked to provide further details on the open-ended survey, common comments were that word cards are quick and easy to make and use. For example, “it is easy to make and doesn’t take long time to make many vocab cards,” “We can study vocabulary when I don’t have enough time,” “It is easy to study vocabulary,” “it is easy to make and it doesn’t take long time to make many vocab cards,” and “I could learn only the words that I don’t know.” In addition, many participants commented on the convenience of word cards, noting that they are easy to carry and can be studied anywhere. Comments included, “I could study them anytime,” “I could study while going to school, taking a train, breaking time of my part time job, whenever I have free time,” “I can learn vocabulary when I am riding on the train. It was also convenient for me to make word cards,” and “we can study wherever we are.”

Moreover, when interviewed, some students mentioned that they prefer word cards over vocabulary notebooks or textbooks such as those they used in junior or senior high school English courses. The main reasons given were that they like the two-sided nature and the ease with which they can be carried or used. The interviewees understood that the two-sided nature of word cards allows for more active and passive recall than a vocabulary notebook where the L1 and L2 forms are placed side-by-side. The interviewees stated that compared to using notebooks, the two-sided nature of the cards not only gave them more choice about how to study, actively or passively, but also increased the effort needed. Two interviewee’s comments sum up the feelings of the others well; “I cannot see the other side so have to try to guess what the word is, it is more difficult but good for me,” and “sometimes I look at the Japanese side and try to remember the English, it is harder … but when no time, reading English and remembering Japanese is better, quicker."

However, because only one participant commented on the intrinsic motivation factor, "motivate up studying vocabulary … It is interesting," the students did not necessarily feel that word cards are an enjoyable or intrinsically motivating activity. This issue is discussed in more detail below.

Learner preferences: premade cards versus self-made. When asked which style of word cards they preferred, 69% (51 respondents) stated that they preferred studying from premade cards such as the ones used in this study, or those available on smartphone applications such as Quizlet. The main reason given was that it is easier and takes less time, but some students also stated that it gives them more time to focus on studying the words.

Some comments concerning the advantages of premade word cards included **"**all is important words. If I make my cards, in it, there are some less important words," "I don't know how to make good words cards," "It is hard for me to find unknown words. I also do not have much time to do that," "It's troublesome for me that making card by myself," and "Student can spend time for the vocabulary practice instead of making vocabulary cards." However, 23 students (31%) stated that they preferred making the cards themselves, the main reason being that they value the productive writing process involved in making the cards. Comments about the writing process included; "It is so important for students to write words," "I can memorize words by writing," "I want to write the vocabulary," "writing is good for memory," and "to write word on my own is easy to remember for me." Overall, when only two choices were provided, more students preferred the premade approach than the self-made.

Learner preferences of the two self-made cards methods. When asked to choose between the two self-made card treatments, one where all the information is provided by the teacher and the other where students find the information themselves, there was no clear preference. Only 44 participants stated a preference, of which 24 (54.5%) wanted to be given the L2 form, L1 meaning, and example sentence from the teacher (student-made no-choice treatment), but 20 students (45.5%) stated that they prefer having an L2 word list and searching for the L1 meaning and sentences independently (student-made with choice treatment). Those that preferred the teacher-led method stated that (a) it was faster an easier, and (b) they trusted the teacher's choice in word selection. For example, representative answers were "it is easy to make and it doesn't take long time to make many vocab cards," and "all words are important if teacher says we need them … good for IELTS improvement." When interviewed, some students who had stated that they preferred premade cards said that their second choice was the teacher-led method for the same reasons; "I don't know how to make good words cards … I don't know best Japanese meaning or good sentence on my own," "It is hard for me to find unknown words and it takes time," and "if teacher gives words I can spend more time studying, less time searching." However, as mentioned, 20 students who stated a preference (45.5%) favored the treatment where they self-select the target items and related information. The main reasons for this preference concerned the autonomous nature of the process, especially the self-selection of target items. For example, typical comments were: “choosing own words is better because students have different words which they don’t understand,” “I could learn only the words that I haven’t know,” and “I could find new meaning when I read the dictionary.” In addition, during the interviews, two students stated that they thought that the teacher chosen L1 meaning was not always the best fit for the L2 target word: “I think that the Japanese word has different meaning to the English …. Japanese word is everyday word, but English one is academic.”

Relationship between word card method and time-on-task (intrinsic motivation). When asked if the type of word card method affected the amount of time spent studying the words during the eight-week experimental phase, the data showed that it had little or no effect. The interviewees said that they spent the same amount of time studying all of the words, regardless of which treatment they came from. Based on observations and interviews, the students combined the cards from all three treatments into one pack or separated them based on their knowledge of the words rather than the creation method. In addition, only one student said that she enjoyed making word cards, and none of the participants stated that either method was more fun or enjoyable than the others. Therefore, the results indicated that none of the word cards methods are intrinsically motivating in general, nor were any of them more intrinsically motivating than the others. However, some interviewees intimated that they felt more ownership with the self-selected and self-made cards (the student-made with choice treatment). For example, one participant summed up the feelings of the interviewees when she stated that “these cards are mine, they are my choice, not teachers.”

Discussion

The results of this study not only highlight a number of key points that contribute to the existing vocabulary literature, but also offer valuable new findings and suggestions relevant to both researchers and teachers. Many of the findings support previous claims in the literature on word cards, and in a broader sense on deliberate learning techniques generally, and other findings add to under-researched areas concerning word cards. However, the main results add new information to the word card literature.

The first main finding of this experiment is that both controlled word card methods resulted in significant short-term vocabulary knowledge gains. After only 15 minutes of making word cards or studying from premade cards, the students demonstrated statistically significant gains on both the immediate posttest active and passive recall measures. Immediately after the initial treatment stage, mean active recall vocabulary scores ranged from six to nine words out of 15 (40-60%) and the participants were able to passively recall between 8 to 10 words, which was 57% and 65% of the target items, respectively. Overall, more than 50% of the 30 target words were learned to some degree after just 30 minutes of time-on-task, thus adding further support to the literature regarding the efficacy and importance of deliberate vocabulary learning in general (Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014; Nation, 2013, pp. 3, 59; Schmitt, 2008). More specifically, the findings support the commonly held belief that word cards are an effective deliberate vocabulary learning technique (Ballance & Cobb, 2020, p. 324-325; Mondria & Mondria-de Vries, 1994; Nation, 2013, p. 437; Reynolds & Shih, 2019). The findings also support those of Experiment 1 and further contribute to the growing body of empirical evidence showing that word cards are a quick and effective deliberate study technique (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014). In addition, the relatively simple nature of the word card methods investigated in this study lend further support to the levels of processing view that even shallow processing can help learners commit new vocabulary knowledge to short-term memory (Craik & Lockhart, 1972, Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Laufer & Rozovski-Roitblat, 2014). Both treatments offered students the chance for multiple encounters with target items, and the opportunity for both active and passive recall or retrieval, adding support for the importance of each process (Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Mondria & Mondria-de Vries, 1994, Nation, 2013, pp. 437–470).

However, surprisingly, the results for research question 2 indicated that the premade word card treatment was significantly better than the self-made treatment at developing short-term vocabulary knowledge. On average, immediately after the initial treatment stage, 40% of the words encountered when making the word cards were actively recalled, but it was 60% for the words encountered in the premade treatment. In addition, while 57% of the words used in the self-made treatment were passively recalled, 65% of the premade word card items were passively recalled. Therefore, in terms of developing short-term vocabulary knowledge, the premade card treatment was more effective. These results indicate that when time-on-task is controlled, and when trying to build short-term vocabulary knowledge, copying words or short sentences is not so beneficial, a finding that supports those of Barcroft (2006, 2007b) and Webb and Piasecki (2018). In the premade treatment, learners only read the words, but they had chance to encounter the cards many more times than in the word card making treatment, so it seems that the number of encounters with a word is more important than depth of processing or having learners complete both receptive (reading) and productive (writing) tasks, a finding that support the assertion made by Folse (2006).

In answer to the third and fourth research questions, the results of the longitudinal part of this study indicated that while all three word card treatments successfully developed longer-term vocabulary knowledge, the student-made with choice treatment was the most effective, but there was almost no difference between the two controlled word card methods (self-made with no choice and premade). The positive learning outcomes offered by all three word card methods offer more support for the long-term memory benefits offered by word card study (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014), and also provide limited support for the Involvement Load Hypothesis (Laufer & Hulstijn, 2001; Laufer & Rozovski-Roitblat, 2014; Pichette et al., 2011). The positive differences between the student-selected with choice treatment and the two word-card methods that gave students no choice in word selection were statistically significant for both dependent vocabulary measures at the delayed posttest stage ($p < .001$ for each comparison). Given that the student-made with choice treatment had the highest task-induced involvement load, and the delayed posttest was the most accurate measure of what learning results from repeated study in usual language learning contexts, it appears that the Involvement Load Hypothesis might be a good predictor of the vocabulary learning outcomes of deliberate study methods. However, the comparisons between the two controlled treatments call this interpretation into question. Not only did the premade card treatment produce significantly better short-term results on both vocabulary measures, the passive recall means were higher for the premade group at the delayed posttest stage. Overall, while the differences were not significant from a pedagogical viewpoint, the premade method, which had the lowest task-induced involvement load, performed slightly better, indicating that the Involvement Load Hypothesis is not a strong predictor of the learning outcomes attributable to the word card methods investigated in this study.

In addition, while not part of the original focus of this study, a further finding of Experiment 2 is that word cards are more effective at developing receptive knowledge than productive knowledge. All three groups at higher passive recall means at each posttest stage than they did active recall. Therefore, a series of paired-sample *t*-tests were conducted comparing the active and passive recall means of each group at both of the posttest stages, the results of which were statistically significant ($p < .005$ for each comparison). These findings add further support to the literature (Behzadi & Haji Pour Nezhad, 2014; Nation, 2013, p. 56; Webb, 2005) and research findings (Komachali & Khodareza, 2012; Kuo & Ho, 2012; Waring, 1997; Webb, 2005, 2009) stating that it is easier to build receptive vocabulary knowledge than productive knowledge.

Finally, the learners had a positive view of using word cards, stating that they are a quick, simple, and efficient deliberate study technique. This finding offers further support for the use of word cards in general (Ballance & Cobb, 2020, p. 324-325; Folse, 2006; Nation, 2013, p. 437–471; Schmitt, 2008) and some survey and interview comments supported the use of word cards over lists or notebooks (Kuo & Ho, 2012; Mondria & Mondria-de Vries, 1994; Nation, 2013, p. 451; Nakata, 2008). The finding that more students preferred the premade word card format than the self-made one, mainly because it is easier and less time consuming support the findings of Dodigovic (2013), and confirm Webb's (2005, 2009) belief that many learners favor the easier method or only make the minimum effort necessary. However, given that the students who preferred making their own word cards valued the writing process, the findings of this study offer affective support for the inclusion of simple copying or writing tasks, a factor that was not take into account in Barcroft's (2004, 2006) negative conclusions regarding such tasks. Finally, based on observations and interviews, none of the three word card methods are intrinsically motivating and did not affect how much time the students spent learning from them over the eight-week experimental period, thus confirming the belief that while effective for vocabulary acquisition, learning from word cards is not necessarily an enjoyable activity (Reynolds & Shih, 2019; Wilkinson, 2017). Therefore, it is important for teachers to introduce additional activities based around word cards that are more intrinsically motivating.

CHAPTER 5

EXPERIMENT 3: LEARNING AND MOTIVATIONAL OUTCOMES OF PEER-TO-PEER VOCABULARY TESTING

Gaps in the Literature

The first gap in the literature is that few researchers have investigated the link between regular testing and motivation for deliberate vocabulary study. Only a few researchers have examined how specific vocabulary activities increase motivation for L2 deliberate vocabulary study (Al-Murtadha, 2014; Thompson, 2014), but none focused on the effects of regular testing, and they mainly relied on self-report data.

In addition, as Webb (2019) observed, few vocabulary researchers have provided qualitative data about individual differences or affective factors. Therefore, a mixed-methods design was used in this study to investigate the extent to which vocabulary testing affects students' extrinsic motivation to carry out deliberate vocabulary study. As a teacher and researcher (Wilkinson, 2015), I have found it difficult to motivate students to devote sufficient time and effort to deliberate vocabulary learning; however, no research currently provides research-based advice to deal with this challenge.

A second gap addressed in this study is that no researchers have offered a convenient and easily implementable test system that enables learners to be reliably tested on individual self-selected vocabulary sets. If students self-select vocabulary items, standardized tests are not idea and it is difficult to provide learners with a method of assessing their vocabulary knowledge and growth. In addition, the inability to offer standardized vocabulary tests for learners who self-select target items leaves teachers with limited ways to assess students on their vocabulary development.

Overall, little research has been carried out investigating the link between vocabulary learning, motivation, and testing (Tseng & Schmitt, 2008), and no studies have been found that provide concrete evidence of the extent to which a test-enhanced learning approach affects motivation for deliberate vocabulary learning.

Purposes of the Study

The first purpose of this study is to introduce a peer-to-peer vocabulary test system and explore its effects on motivation for deliberate vocabulary study. The peer-to-peer vocabulary testing system under investigation has previously received favorable feedback from university students (Wilkinson, 2015), but more rigorous research is needed.

The second purpose is to to investigate whether the peer-to-peer test system accurately measures learners' knowledge of the target vocabulary items. As the tests are administered by learners, the scores are potentially unreliable as students might be overly helpful during the test or too generous when awarding points. If the test system proves reliable, it will provide teachers with a tool to assess students' individual vocabulary knowledge. However, if the test is motivating yet unreliable, while it can be viewed as a useful affective activity that can serve as a method for providing feedback to students on their vocabulary learning progress, it should not be used for assigning grades.

Overall, the experiments and analyses carried out in Experiment 3 aim to explore the link between student-centered vocabulary learning, testing, and motivation.

Research Questions

The answers to the research questions below offer a range of information regarding the effects of repeated peer-to-peer vocabulary testing in a foreign language course. Answers to the questions shed light on the system's effect on learners' motivation for deliberate vocabulary study and assess the reliability of the student-to-student test system in order to investigate if it can be used as part of course assessment.

1. To what extent does the test system affect extrinsic or intrinsic motivation?
2. How is the peer-to-peer test perceived by learners and how does this affect the answer to research question 1?
3. How reliable are students as examiners?

Methods

Participants

The series of experiments conducted in this chapter were conducted at the same private women's university described in Experiment 1 and 2, but they were carried out in the following academic year with different courses and different participants.

When analysing the data for the first two research questions, there were originally 86 first- to fourth-year university students, but one was removed for missing one of the experimental stages; thus, there were 85 female Japanese participants, ranging in age from 18 to 22. However, when answering research questions 3, as described in more detail later, complete data were collected for only 50 of these participants, so $N = 50$ for the last experiment described in this chapter. The details of the four courses and the participants in each course are described below.

Critical reading and discussion (CRD). The Critical Reading and Discussion course was an elective English course open to students from any department in the university in their second to fourth years of study on an undergraduate degree program. There were 24 female participants enrolled in this course from five departments; Language Sciences, English, Philosophy, International Relations, and Sociology. The participants were aged between 19 and 21 (M = 19.6 years). The course met once a week for 90 minutes and was focused on reading news articles in English, writing multiple summary and response reports, giving informal presentations, and leading discussions based on the reports. In addition, there was a deliberate vocabulary study component included on the official course description and syllabus. This component involved the students self-selecting approximately 20 words each week from the NGSL and NAWL word lists, and the deliberately studying them outside of class using the Quizlet application (more details are provided regarding this later on in this chapter). In addition, some class time was devoted to additional activities that promote the learning or review of the target words, for example, pair discussions about the target items, and trying to use the words in sentences. The students also engaged in speed-reading using materials available at https://www.wgtn.ac.nz/lals/about/staff/paul-nation#Sonia. Based on in-class observations and homework assignments, approximately 25% were viewed as being highly motivated; these students had zero or one absence, had completed all the homework to the required level, and actively participated in each class. When questioned, these students stated that they enjoyed the course activities and did the homework to the best of their ability in order to maximise their learning. However, approximately 55% of the participants can be viewed as having medium to low motivation as they were absent more than once and did not complete at least on aspect of the required homework fully on more than one occasion. In addition, 20% were absent three times, the maximum allowable in order to pass the course, they did not complete the required homework multiple times, and they submitted incomplete files at the end of the course. In sum, the participants had mixed levels of intrinsic and extrinsic motivation. 16 students had taken the TOEIC tests and scores ranged from 470 to 660 (M = 520), so they were intermediate proficiency learners. Based on the pre-course VLT administration, the participants had an average of 80% knowledge of the first 2,000 high-frequency words, 70% of the third 1,000, and less than 50% knowledge of the words in the Academic Word List.

Advanced speaking. The Advanced Speaking course was another elective English course open to students from any department of the university in their second to fourth years of study. The course met once a week for 90 minutes and was focused on researching controversial topics in order to carry out lengthy discussions and debates. In addition, the same deliberate vocabulary study component as described in the CRD participant section was part of this course. The participants in this course were between 19 and 21 years of age (M = 19.2 years). There were 16 female participants who came from four departments; Language Sciences, Philosophy, International Relations, and

Sociology. Because this was an elective course, the students appeared to be intrinsically or extrinsically motivated to study English. Based on in-class observations and homework assignments, this generally seemed to be the case. For example, 70% of the students had zero or one absence, had completed all the homework to the required level, and actively participated in each class. When questioned, these students stated that they enjoyed the course activities and did the homework to the best of their ability in order to maximise their learning. However, approximately 30% of the participants were absent two or three times and did not complete the required homework fully on one or two occasions. All students had taken the TOEIC test and scores ranged from 480 to 690 (M = 545). In terms of vocabulary knowledge, on average the participants had 85% knowledge of the first 2,000 high-frequency words, 70% of the third 1,000, and less than 60% knowledge of the words in the Academic Word List. Therefore, given the similar TOEIC and VLT score range, the participants were at a similar proficiency level to the participants in the class described above.

First-year English seminar (FYES). The First-Year English Seminar was a required course for first-year students in the English department. The course met twice a week for a total of 180 minutes and was focused mainly on academic writing and IELTS preparation. As with the other courses described above, there was a deliberate vocabulary study component in the course. The students took the IELTS test at the end of the fall semester and they planned on studying abroad in their fourth, and sometimes fifth, semester. The 24 female Japanese students in this course were 18 years old and their

TOEIC scores ranged from 280-370 (*M* = 310), and they were placed in this class based on these scores. The VLT indicated that the participants had less than 70% knowledge of the first 2,000 high-frequency words, less than 40% of the third 1,000, and less than 30% knowledge of the words in the Academic Word List. There were seven FYES classes; one advanced, four intermediate, and two pre-intermediates; this class was one of the pre-intermediate groups. Based on observations and assignment checking, the students had mixed levels of motivation and their individual motivational levels changed throughout the course. 30% of students were highly motivated, attending all or most classes, actively participating in most classes, and completing all or most assignments fully. The other 70% were absent between three and six times, did not complete some assignments, and did not participate actively in some classes.

Second-year English seminar (SYES). The Second-Year English Seminar was a required course for second-year students in the English department. The course met once a week for 90 minutes, it was focused on academic writing, and there was a deliberate vocabulary study component. There were 21 female Japanese students who were 19 or 20 years old (*M* = 19.4). Their latest IELTS test scores ranged from 5 to 6.5 (*M* = 5.5), but the class was not streamed based on English level. Based on the pre-course VLT administration, the participants had an average of 70% knowledge of the first 2,000 high-frequency words, 60% of the third 1,000, and less than 50% knowledge of the words in the Academic Word List.

This class was one of seven SYES classes. Based on observations and assignment checking, the students had mixed levels of motivation and their individual motivation levels changed throughout the course. 20% of the students were highly motivated, attending all or most classes, actively participating in most classes, and completing all or most assignments fully. The other 80% were absent one to three times, did not complete some assignments, and did not participate actively in some classes.

Research Design

A fixed mixed-methods design was used as the mixed-method design was decided at the outset of the study (Creswell & Clark, 2017, p. 52) (See Figure 3).

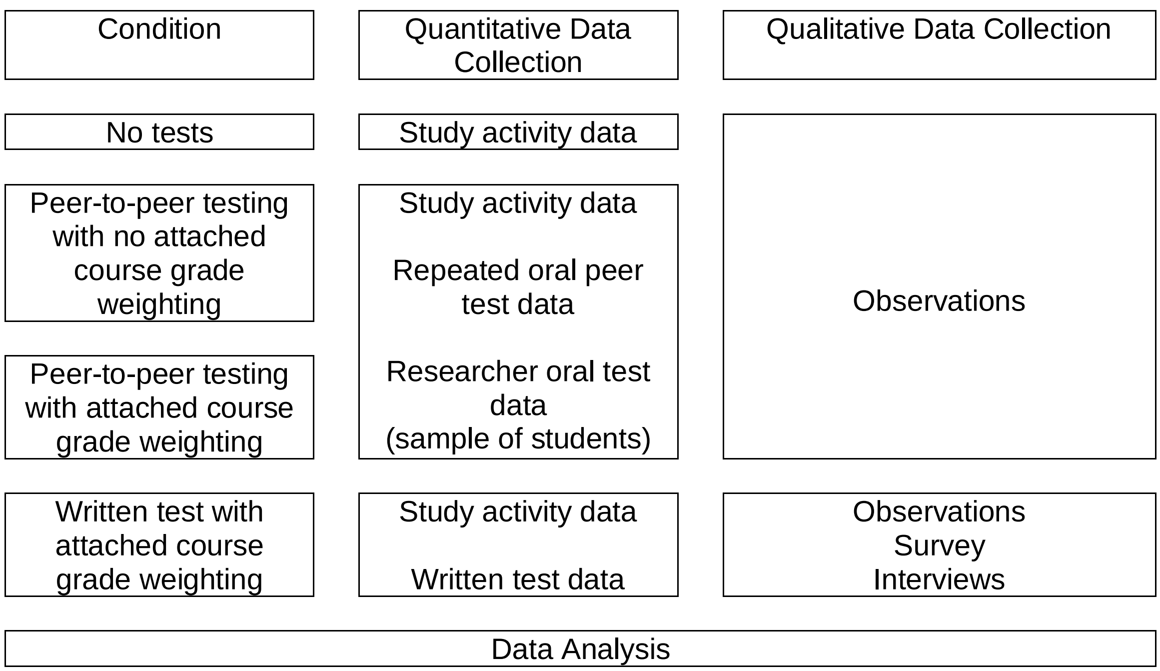

Figure 3. Research design for Experiment 3.

Creswell (2014) described this design as a convergent mixed methods design because both forms of data were collected throughout the study (p. 15). However, it was primarily a quantitative study with additional qualitative data collected and analyzed to inform the quantitative results. In addition, the design includes longitudinal data regarding motivational effects and the reliability of students as examiners.

Materials

Vocabulary items. The vocabulary items used in the study were items the learners had self-selected from the NGSL and NAWL word sets. Based on their VLT scores, the students were provided with the second and third 1,000 word-level of the NGSL and the complete NAWL. The students had the choice of studying from premade cards or to make their own word cards, but almost none of the students chose to make their own word cards. Therefore, the vocabulary materials used in this study were the pre-made sets provided on the Quizlet application described below. The application provided the L2 target word along with an L1 meaning. The students were required to study from vocabulary sets by self-selecting approximately 30 words per week for 12 weeks as part of the course requirements.

Deliberate vocabulary learning application: Quizlet. The Quizlet application is designed for deliberate vocabulary learning. The application was set up with three vocabulary packs; two based on the NGSL and one from the NAWL. The only information provided in the application was the L1 form of the target items and an L2 meaning. The application offers five study activities, but these were adjusted in order to improve the research design by better controlling for time-on-task. The five study modes and the specific adjustments made to the application settings are described below.

1. ***Learn.*** This mode allows for both active and passive recall practice and utilizes three learning methods; word card study, multiple-choice options about the meaning of a target word, and writing the meaning of the word based on the L1 or L2 form provided. Normally, students can choose to engage only in passive recall, only active recall, or a combination of both, and can select one to three of the practice methods. However, for this study, the students were instructed to use only the multiple-choice and flashcard modes and were told to select the option that allowed them to study both passively and actively. The setting also allowed the students to choose how many words to study in a session, but this was set to 50 words. To do this, time in class was devoted to showing the students how to set the target word *pack* size to 50. While it was not totally possible to control this; students could change it if they wish, this was checked throughout the course during the various vocabulary activities and when administering the teacher-administered verbal test.
2. ***Flashcards.*** This electronic flashcard application allows for both active and passive recall practice, but not a combination of the two in the same sitting. The settings can be changed so that when learning the cards, either the L1 form of the word is shown and the students must recall the L2 form, or vice versa. After seeing the first side of the flashcard, students tap the card to show the answer on the other side of the card. For this study, the students were encouraged to regularly change the settings to allow

them to study both actively and passively. The students were encouraged to regularly change the settings to allow them to study both actively and passively, and the number of words to include in a study session was set to 50 words.

3. ***Write.*** This written mode allows for both active and passive written output practice, but not a combination of the two in the same study session. Students can choose either active or passive mode but not a combination of both, so the L1 or L2 form of a word is shown and students write the corresponding meaning. Because many students aimed to take exams such as the IELTS or TOEFL iBT, both of which require writing, and in order to control time-on-task, the students were encouraged to study mainly actively. In addition, due to the more time-consuming nature of the writing task, the number of target items was set to 20.
4. ***Match.*** This mode displays the L1 and L2 forms of a preselected number of target items and students must match the L1 word with its L2 meaning or vice versa. The activity is timed, and students are told when they achieve a personal best. The setting also allows students to choose how many words to study in a session, and this was set to 50 words.
5. ***Test.*** This mode provides three test question types; true/false, multiple-choice, and write, but selects only 20 words per test. Words are recalled or written both actively and passively. Students are told their current and highest test score each time. No modifications were made to this other than the students were told to include all question types.

The quizlet application allows the teacher to observe students' study practices, and these observations were used when designing the dependent motivation measures. A small pilot study was carried out with six students to investigate how long one study session takes for each of the five activity types. When set up as described above, by modifying the application settings, each of the five study modes took approximately five minutes to complete, more details regarding this issue are provided in the next section.

Instruments

Measures of motivation. The data collection method allowed motivation to be measured in two ways under each condition, each measuring related but distinct study habits. The two measures of motivation were (a) the number of times students chose to study vocabulary deliberately and (b) the number of activities the students completed.

First, as the most basic measure of extrinsic motivation, the number of times students chose to study vocabulary deliberately under each condition was used. These data were collected by monitoring the number of students who chose to do some form of vocabulary study each week under the four test condition stages, each of which lasted three weeks. Unfortunately, the Quizlet application does not provide data regarding the length of time students spend studying the words. When collecting these data, the students were monitored each week, and any student who completed at least one of the Quizlet study activities in that week was awarded 1 point. Therefore, under this measure, during each three-week test condition stage each participant could obtain a score of between 0 and 3 points.

Table 25 provides more detail about how this measure was recorded and interpreted.

Table 25. *Description of Measure 1: Study Counts*

Score	Participants action	Level of Motivation
0	Did not study at all during the three weeks	No motivation
1	Studied only one week out of the three	Low motivation
2	Studied two weeks out of three	Medium motivation
3	Studied every week	High motivation

These data provide basic but distinct information regarding the students' study choices. The basic study count indicates whether the test conditions resulted in more students doing some form of vocabulary study under each condition. However, it also provides a basic measure of repetition and the extent to which the participants spaced their learning. For example, if a participant did not study for two weeks, even if they did many activities in the third week, they would get a score of 1 point.

The second measure of motivation was the number of activities the students completed during each test condition stage; this is likely the most accurate measure because it most closely represents time-on-task and repetitions. The data were collected by counting and totalling how many of each of the five types of study modes the students engaged in during each test condition stage. As discussed above, by managing and setting up the Quizlet settings, all activities were assumed to take a similar amount of time. For example, the number of target items in each mode and the types of tasks in each activity were modified, and only complete activities were counted.

To check if this design resulted in each activity taking a similar length of time, a small pilot study with six students was carried out to measure the time it took to complete each of the five Quizlet study activities. All six students completed two study sessions using each of the five study activities modes—learn, flashcards, write, match, and test—and their times were recorded. In addition, after the study was completed, I asked some participants about how long each activity took. Table 26 displays (a) details about the activities as set up in this study and (b) the approximate times based on the pilot study and post-treatment interviews.

Table 26. *Details About the Activities Set-up and Time-on-task Estimates*

Quizlet activity	Study mode	Number of target items (app settings)	Approximate time to complete
Learn	Multiple Choice (MC) and flashcards	50 words	4 to 5 minutes
Flashcards	L1 $\rightarrow$ L2 and L2 $\rightarrow$ L1	50 words	4 to 5 minutes
Write	L1 $\rightarrow$ L2	20 words	6 to 7 minutes
Match	Match Form and Meaning	50 Words	4 to 5 minutes
Test	True/False, MC, Write	20 Words	4 to 6 minutes

Note. The times are based on the minimum and maximums observed during the pilot study and times reported by students.

Peer-to-peer tests. The two peer-to-peer tests acted as both treatment conditions and measurement instruments. First, the two tests acted as part of the treatment conditions under investigation in research questions 1 and 2, but they were also used as the measurement instruments to answer research question 3. The tests were based around Nation's (2013) description of word knowledge. While not all aspects are included in the tests, they do include both passive and active recall, spelling, pronunciation, and generative use in a sentence.

Two test types were used, a verbal test and written test. In the verbal test, the learners exchanged their mobile phones with the Quizlet application open and then asking their partner various questions about randomly selected target items. The verbal test took 20 minutes, and all learners were tested on 20 words. They were awarded 1 point if they (a) actively or passively recalled the correct meaning of a word, (b) correctly spelled the L2 word verbally or by writing it, and (c) answered another question chosen by their partner. If teachers so wish, the tests can be adapted so students can gain extra points by successfully making example sentences, however, as it is difficult to ensure or check that the students correctly judge the sentences, this test feature was not included in this study.

The second student-generated test was a written test in which the learners randomly selected a specified sample of their partner's target words and copied either the L1 or L2 form of the word onto the test sheet. The participants opened Quizlet, exchanged smartphones, selected 10 words, entered the L1 forms onto the test sheet, and then selected a further 10 words and entered the L2 form onto the test sheet. As part of the test, the learners then wrote the L1 or L2 meaning, the part of speech, and showed knowledge of other words belonging to the same word family (see Appendix H for the peer-to-peer verbal test and Appendix I for the student-generated written test).

Procedures

To answer the first research question, quantitative data regarding the effect of the different test conditions came mainly from observations of the participants' autonomous vocabulary learning habits. However, classroom-based observations and informal interview data were collected and analyzed in order to triangulate the quantitative data.

To assess the motivational effects of the test conditions, data were collected over the course of the experiment by digitally monitoring students' study choices using the *Class Progress* feature available on the teacher version of QUIZLET. As the participants were unaware that the teacher had access to this data, the observation procedure did not influence the students' decisions.

During the experimental stage, there were four treatment conditions, each lasting three weeks, so the total length of the experiment was 12 weeks. The conditions were: (a) no vocabulary testing, (b) regular peer-to-peer vocabulary testing that did not count towards the course grade, (c) regular peer-to-peer vocabulary testing that the participants believed counted toward the course grade, and (d) pre-announced written vocabulary tests that the students believed counted towards the final grade.

The order of the four testing conditions was deigned to give participants the maximum opportunity to display preexisting motivation levels for deliberate vocabulary study and follows a continuum of self-regulated motivation. For example, any deliberate vocabulary study undertaken during condition one (no testing) would be a result of participants own intrinsic or extrinsic motivation levels, they would be choosing to study because they either enjoy doing so or have some personal extrinsically motivating factor.

Likewise, under the second condition, as there was no course grade attached to the weekly verbal test scores, any additional study time would be a result of the introduction of the test; for some reason (identified through the analysis of the qualitative data), the test motivated them to study more. Then, the third test condition, while counting towards the final course grade, does not resemble the type of formal testing that participants are used to and could still be found to be intrinsically motivating, and was therefore likely to induce less anxiety or extrinsic threat. Finally, the last testing stage most closely resembles the formal testing that participants are used to and cannot be assumed to be intrinsically motivating. More details about procedures and conditions involved in the experiment are described below.

First, at the beginning of the course, learners were instructed to study using the Quizlet application described above. They were instructed to study vocabulary throughout the course, and they had the option to make their own vocabulary materials and/or study from premade materials, both of which were based on the NGSL and NAWL. The participants were told to study as many words as they could, but the minimum was set at 30 words per week. At the start of the study, the participants were told that 20% of their course grade would come from the deliberate vocabulary learning component. As with most of my courses, vocabulary development was listed on the syllabus; although the students knew that it accounted for 20% of their course grade, no specific details were mentioned about grading; they were told to make and study vocabulary materials using Quizlet. In addition, as would have been done if the experiment was not taking place, to provide further motivation for studying vocabulary, repeated explanations about the necessity of vocabulary and the benefits of deliberate vocabulary study were given across all treatment conditions.

In the first condition (no tests), during the initial three weeks of the course, no vocabulary testing took place, so this period acted as a control. However, during this three-week period, the benefits of deliberate study from word cards were repeatedly emphasized, and I repeatedly encouraged the students to study deliberately vocabulary using the Quizlet application. In addition, the students were engaged in various vocabulary-focused activities in each class for approximately 15 to 20 minutes. These activities included teaching the meaning of their target vocabulary items to partners, using their words in sentences or conversations, and consulting the teacher for additional information such as other word forms, pronunciation, or example sentences. The learners outside-of-class study choices were monitored each week using the progress monitoring feature of the Quizlet application, but the learners were unaware of this. In addition, the learners were observed for five or 10 minutes before class, for those who arrived early, to see what activities, if any, they were engaged in.

After the initial three weeks, the second treatment stage involving regular peer-to-peer vocabulary testing not affecting the course grade began. During this stage, the students were tested on their vocabulary using the spoken peer-to-peer test once a week for three weeks. For classes that met once a week, the Stage 1 vocabulary activities were largely replaced with the test, but for the twice-a-week classes, one class focused on the test and the other included the same kind of vocabulary activities as in Stage 1. The test, which took 20 minutes to complete, offered ample chances for revision, production, and self-assessment, practices that are the main focus of the vocabulary activities used in Stage 1. Under this test condition, the students were not encouraged to study their word cards as much as in Stage 1, but they were reminded that they should be doing so. In this stage, the students were told that the test scores would not be used when calculating their final grade. During Stage 2, the students' out-of-class vocabulary study choices were again observed using the progress checking feature of Quizlet, and observations were carried out in class five minutes before the start of each lesson.

Next, Stage 3, the regular peer-to-peer vocabulary testing affecting the course grade condition, followed the same procedure as Stage 2, with one notable difference; the students were informed that their average scores for the tests taken during this three-week period would be graded and the results used when calculating their final course grade. No details were given about how much of the 20% vocabulary grade would come from the tests, just that the results would significantly affect their course grade.

The fourth treatment stage, written vocabulary tests affecting course grade, measured the effects of introducing more formal written tests similar to what the students had experienced in their previous language courses. The students were informed that they would take at least two written vocabulary tests aimed at assessing their individual vocabulary knowledge at some point in the preceding three weeks. The students were informed that the test results would also count towards their final course grade. After placing the participants into pairs and having each student randomly select 20 words from their partner's study materials—10 items to be tested actively (L1-L2) and 10 items to be tested passively (L2-L1)—the participants took the test. After the test, the participants scored their partner's test using the information in their original vocabulary materials.

To answer research question 2, in-class observations, survey data, and interview data were collected and analyzed to assess the participants' perceptions of the different test conditions. Based on advice by Creswell (2014, pp. 200–209), Creswell and Clark (2017, pp. 54–55), and Hatch (2002, pp. 93–133), observations, questionnaires, and interviews were used to collect and triangulate the quantitative and qualitative data. The survey (see Appendix J for the survey) included both Likert scale items and open-ended questions. The questionnaire was written in English, but students had the choice of answering the open-ended questions in Japanese. In addition, observations were carried out during the in-class vocabulary activities and testing stages. The observational data aimed to provide insights into what the students did during the tests such as the choice of passive and active recall testing and what word knowledge questions the students focused on. The observations also offered insights into the class atmosphere during the test, thus helping to assess the participants' perceptions of the test. As noted by Creswell (2014, pp. 186–189), Creswell and Clark (2017, pp. 178–180), and Hatch (2002, pp. 73–78), observational data often highlights a variety of issues that the researcher had previously not considered, and the data gleaned from the observations can be used for the selection of participants for follow-up interviews. The final data collection method involved semi-structured interviews, but due to time and logistic constraints, only a random sample of the participants were chosen for interviews. The interviews were carried out in English, but there was access to a translator if the students felt the need to express their ideas in

Japanese. The interviews were carried out following the guidelines set out by Hatch (2002, pp. 91–116). For example, an interview protocol was drawn up, but I was willing to create new questions based on the individual interviewee's responses, interview notes were taken, and interviews were recorded unless the participants did not feel comfortable being recorded.

Finally, in order to answer research question 3, the students' test scores were collected. All data were collected during one academic semester with the same intact groups. Three types of test data were collected: student-to-student verbal test scores, teacher-to-student verbal test scores, and student-generated written test scores. To obtain these data, after being tested using the verbal peer-to-peer test system once a week for approximately six weeks, an average test score was calculated for each student. In addition, a random selection of students was tested by me once during the study period using the verbal test system. The last data type came from the written test administered to all students in the last test condition stage. While all tests were compared, the scores on the teacher-administered verbal test and the student-generated written test were deemed to be accurate measures of the students' vocabulary knowledge, and both were compared to the student-to-student verbal test data to help answer the third research question.

In summary, the data needed to answer the research questions were collected over a 12-week period and the use of the quantitative and qualitative data collection methods described above aim to offer holistic and triangulated data regarding the motivational effects of regular vocabulary testing, and the reliability of the test system.

Data Analysis

Motivational effects of different test conditions. The main constructs under examination in research questions 1 and 2 are general perceptions of the test conditions and their effects on extrinsic motivation. In order to assess the motivational effects of the four test conditions, analyses were carried out using three Friedman tests in IBM SPSS Statistics 26, one for each dependent measure of motivation for deliberate vocabulary study. The Friedman test is a non-parametric alternative to the one-way repeated-measures ANOVA, and as such, is sometimes referred to as the Friedman's ANOVA. The test is commonly used when the groups are related; they contain the same participants and each group is measured repeatedly on the same dependent variable.

The Friedman test was chosen over repeated-measures MANOVA or one-way repeated-measures ANOVA because the assumption of normality was violated, and the dependent variables were measured on an ordinal scale. The Friedman test is suitable when ANOVA or MANOVA assumptions have been violated (Field, pp. 2018, 321) and for designs using ordinal-scale measures (Laerd Statistics, 2019). The independent variable was test condition (four levels: no tests, regular peer-to-peer vocabulary testing not affecting course grade, regular peer-to-peer vocabulary testing affecting course grade, written vocabulary tests affecting course grade), and the dependent variable was one of the two measures of motivation: number of times students studied and number of learning activities completed.

Additional data collected during the study from observations, surveys, and interviews were used to triangulate the quantitative data. As mentioned above, the observational data gained from the Quizlet progress checking did not provide details about time-on-task, just how many times the students engaged in a particular activity; therefore, the analysis of this extra data sheds further light on issues that might have affected the statistical results.

Students' perceptions of the different test conditions. Qualitative data from the observations, surveys, and interviews were analyzed to provide insights into the learners' perceptions of the test conditions and how these were related to changes in motivational levels over the course of the study. Observations, interviews, and survey data were analyzed qualitatively and quantitively to examine the participants' perceptions of each test condition and investigate whether either was viewed more favorably than another. The questionnaire included a combination of Likert scale questions and open-ended questions. The questionnaire was written in English, but the participants could answer in their L1 (Japanese) if they wished. The questions focused on how the learners felt about each test condition and test type. The interviews were conducted to garner more detailed information about the same issues.

Reliability of students as test administrators. In order to answer research question 3, three types of test data were analyzed; student-student verbal test scores, teacher-administered verbal test scores, and student-generated written test scores. The data were analysed using four types of statistical analyses; Friedman's test, Kendall's coefficient of concordance, Spearman's correlation, and intraclass correlation coefficient analysis.

The Friedman test analysis was used to investigate if there were significant differences between the distributions of three related groups' test mean scores. The Kendall's coefficient of concordance analysis was run to investigate the extent of agreement between the three tests administered to each participant. As the two types of verbal tests—student-to-student verbal test and teacher-to-student verbal test—represent the same test administered by different raters, and the student-generated and scored written test used the same question types but was administered and scored by a different rater, the analysis was suitable to assess agreement between the different tests. The Spearman's rank-order correlation analysis was used to measure the strength and direction of the association between the test scores. This involved running one analysis for each of the three test pairs. Finally, intraclass correlation coefficient analysis (ICC), which is an index of the reliability of the ratings for a single judge or measure, was carried out. In this study, the ICC analysis was used to investigate the reliability of the ratings produced for each participant by the three vocabulary test measures.

By running the four types of analyses, it is possible to evaluate the reliability of the three tests and to assess the reliability of students as examiners, thus determining if the student-generated tests can be used in place of more standardized teacher-generated written tests; tests that cannot be used when students are self-selecting their target items.

Assumptions underlying the statistical tests. Two assumptions must be met with the Friedman test; there must be one dependent variable that is measured at the continuous or ordinal level, and one independent variable that consists of three or more categorical, related groups, or matched cases (Laerd Statistics, 2019). Both assumptions were met for all of the Freidman tests analysis carried out in this study. First, when assessing motivational effects, for each Freidman test, there was one dependent ordinal measure (motivation) and one independent variable (condition) containing four related groups; all participants experienced all conditions. Second when assessing the reliability of the tests, the assumptions were met as there was one dependent continuous measure (vocabulary knowledge) and one independent variable containing three related groups (student-student verbal test, teacher-administered verbal test, and student-generated written test). However, with regard to the dependent measure, while all three tests used the same vocabulary knowledge question types and dependent measurement scale, the tests were not the same. For example, all tests measured vocabulary on a continuous scale ranging from 0 to 20, and the two verbal tests were the same, only administered by different people (raters), but the written test format was different. The written test used the same question types as the verbal test (active and passive recall, spelling, and word family knowledge), but the format and mode of delivery were different; written instead of spoken questions and answers. Therefore, the results should be interpreted with caution, and the other three statistical tests were carried out to triangulate the results of the Friedman test and provide more confidence in the conclusions.

The assumptions of the Kendall’s coefficient of concordance analysis were met as (a) the judgements made by the raters (test scores) were measured on a continuous scale, (b) each test and rater was assessing the same participants, and (c) the three raters were independent; one rater’s judgement was not able to affect the other rater's judgement (Laerd Statistics, 2019).

The assumptions underlying the Spearman’s rank-order correlation analyses were met because each analysis involved (a) two variables measured on a continuous scale, (b) the two variables represent paired observations where each paired observation reflects the score on each variable for a single participant), and (c) there was a monotonic positive relationship between the two variables. This was checked by producing and visually inspecting a scatterplot.

The major requirement of intraclass correlation coefficient analysis is there must be the same number of ratings (scores) for every case rated (participant) and this requirement was met. Another issue that affects the set-up choices made in SPSS is whether there are consistent raters for all ratees, for example, do the same three raters assess every participant? In this study, the answer is no, so this analysis was a one-way random model.

In summary, all of the assumptions and requirement underlying the statistical analysis were checked and met, thus the results discussed below can be read with confidence. In addition, where possible, the survey, interview, and in-class observation data were used to provide further insights into what learner factors might have influenced the statistical results.

Results

Motivational Effects of Different Test Conditions

Analyses were carried out using three Friedman tests. The independent variable was test condition (four levels: no tests, regular peer-to-peer vocabulary testing not affecting course grade, regular peer-to-peer vocabulary testing affecting course grade, peer-generated written vocabulary test affecting course grade). The dependent measure was one of the two measures of motivation: (a) number of times students studied and (b) the number of learning activities completed under test each condition.

Number of times students studied (study sessions) under each condition. The descriptive statistics for this analysis are shown in Table 27. As can be seen in Table 27, the mean and median values were lowest in the no-test condition and were highest in the peer-to-peer vocabulary testing not affecting course grade condition.

Table 27. Descriptive Statistics for the Four Treatment Conditions on the First Dependent Measure: Study Sessions

	No tests	Peer-to-peer vocabulary testing not affecting course grade	Peer-to-peer vocabulary testing affecting course grade	Peer-generated written vocabulary test affecting course grade
M	1.33	2.35	2.06	2.13
Min.	0	0	0	0
Max.	3	3	3	3
SD	.88	.81	.98	.83
SE	.09	.09	.11	.09
Mdn	1.00	3.00	2.00	2.00
95% CI (M)	[1.14, 1.52]	[2.18, 2.52]	[1.85, 2.27]	[1.95, 2.31]
95% CI (Med.)	[1.00, 2.00]	[2.00, 3.00]	[2.00, 3.00]	[2.00, 2.00]

Note. *N* = 85. CI (M) = Confidence interval of mean. CI (Med) = Confidence interval of median.

These data indicate that on average, when there were no vocabulary tests, students chose to deliberately study their target vocabulary one week out of three ($M = 1.33$). In contrast, for all conditions that included vocabulary testing, the students studied in at least two weeks out of three on average. However, of the three conditions in which the students were tested, they studied most when the scores did not affect their course grade. In addition, the minimum statistics indicate that some students did not study under any condition, and the raw data indicate that some students were consistent in their study habits under each condition.

A Friedman test was carried out in order to investigate whether the positive changes and differences in motivation levels under the various conditions were statistically significant. The independent variable was test condition (four levels: no tests, regular peer-to-peer vocabulary testing not affecting course grade, regular peer-to-peer vocabulary testing affecting course grade, written vocabulary tests affecting course grade), and the dependent motivation measure was the number of times the students studied. There was a significant difference among the four levels of motivation as assessed by the test conditions, $\chi^2(3) = 83.80$, $p < .001$.

Next pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at $< .012$), and the Wilcoxon signed Rank Test Statistic (r) was used to calculate effect sizes. According to Field (2018, p. 328) and Laerd Statistics (2019), $r = .10$ is considered small effect size, $r = .30$ represents a medium' effect size, and $r = .50$ or above indicates a large effect size). This analysis revealed statistically significant differences between the no-test condition and all other testing conditions and medium to large effect sizes.

Specifically, the analyses revealed statistically significant increases in motivation levels from the no-testing condition (Mdn = 1.00) to the regular peer-to-peer vocabulary testing not affecting course grade condition (Mdn = 3.00), $r = .52$, $p < .001$, regular peer-to-peer vocabulary testing affecting course grade conditions (Mdn = 2.00), $r = .43$, $p < .001$, and written vocabulary tests affecting course grade conditions (Mdn = 2.00), $r = .54$, $p < .001$. These results indicate that compared to having no vocabulary tests, all three testing conditions significantly increased the number of times the students studied, however, further pairwise comparisons with a Bonferroni correction for multiple comparisons indicated that the differences between the three conditions where students were tested on their vocabulary knowledge were not statistically significant, ($p > .12$).

The analyses indicate that all forms of vocabulary testing used in this study increased motivation as measured by the number of times the students studied in each condition. However, whether or not the tests are used for grading, or whether the tests take an oral or written form seems to have little effect.

Number of activities completed. Before discussing the results of the final Friedman test analysis, the descriptive statistics are shown in Table 28. As can be seen, the mean and median values were again lowest in the no-test condition and highest in the peer-to-peer vocabulary testing not affecting course grade condition. When there were no vocabulary tests, the students chose to complete only 5.54 deliberate study activities in three weeks. In contrast, under the peer-to-peer vocabulary testing not affecting course grade condition, the students completed over 17 activities. As this figure decreased to 10.61 for the verbal test that affected the grade condition and 12.90 for the written test condition, the results again indicate that of the three conditions in which students were tested, they studied the most when the scores did not affect their course grade. However, the minimum and maximum statistics show that some students did not study at all under any of the conditions, while others studied a great deal. In addition, the analysis of the raw data indicate that some students studied in a consistent manner under all conditions.

Table 28. Descriptive Statistics for the Four Treatment Condition on the Second Dependent Measure: Number of Activities Completed

	No tests	Peer-to-peer vocabulary testing not affecting course grade	Peer-to-peer vocabulary testing affecting course grade	Peer-generated written vocabulary test affecting course grade
M	5.54	17.17	10.61	12.90
Min.	0	0	0	0
Max.	33	142	121	122
SD	7.42	24.55	16.48	20.36
SE	.81	2.66	1.79	2.21
Mdn	3.00	8.00	5.00	6.00
95% CI (M)	[3.94, 7.14]	[11.88, 22.47]	[7.06, 14.17]	[8.51, 17.29]
95% CI (Med.)	[2.00, 5.00]	[5.00, 14.00]	[4.00, 7.00]	[5.00, 8.00]

Note. *N* = 85. CI (M) = Confidence interval of the mean. CI (Med) = Confidence interval of the median.

Next, a Friedman test was carried out to investigate whether these differences were statistically significant. The independent variable was test condition (four levels: no tests, regular peer-to-peer vocabulary testing not affecting course grade, regular peer-to-peer vocabulary testing affecting course grade, written vocabulary tests affecting course grade), and the dependent motivation measure was the number of activities students completed. The level of motivation differed significantly for the different test conditions, $\chi^2(3) = 72.278$, $p < .001$.

To investigate where these differences exist, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons (p is significant at < .012), and the Wilcoxon signed Rank Test Statistic (r) was used to calculate effect sizes. This analysis revealed significant differences between the no-test condition and all other testing conditions and medium effect sizes.

Specifically, the analyses revealed statistically significant increases in motivation levels from the no testing condition (*Mdn* = 3.00) to the regular peer-to-peer vocabulary testing not affecting course grade condition (*Mdn* = 8.00), $r = .48 p < .001$, regular peer-to-peer vocabulary testing affecting course grade conditions (*Mdn* = 5.00), $r = .31$, $p < .001$, and written vocabulary tests affecting course grade conditions (*Mdn* = 6.00), $r = .39$, $p < .001$. The results of this analysis and the descriptive statistics indicated that all three testing conditions significantly increased the number of activities students completed compared to having no vocabulary tests. Further pairwise comparisons with a Bonferroni correction for multiple comparisons indicated that the differences between the three conditions in which the students were tested on their vocabulary knowledge did not differ significantly ($p > .12$).

The analyses indicate that all forms of vocabulary testing used in this study increased motivation as measured by the number of activities the students completed. However, whether or not the tests are used for grading, or whether the tests take an oral or written form seems to have little effect.

In summary, the results revealed statistically significant increases ($p < .001$) in both measures of motivation levels from the no-test condition to all three conditions in which students were tested; thus, all forms of peer-generated vocabulary testing increased motivation for deliberate vocabulary study. This indicates that the tests encouraged the students to study more often and to carry out more deliberate study activities. However, there were no significant differences between the three test conditions on any of the dependent motivation measures. As discussed below, this finding indicates that assigning grades to the student-generated tests does not positively affect motivation levels, nor does administering the test in a traditional paper-based form.

Participants' Perceptions of the Test Conditions

Survey, observation, and interview data analyses were carried out to investigate how the tests and test conditions were perceived by the participants, and how their perceptions might have influenced the statistical results discussed above.

First, the survey, observation, and interview data analysis show that the students were least motivated in the no-test condition and preferred the other test conditions, a result that supports the statistical results discussed above. During the three weeks when the students were not being tested (Condition 1), few students were seen studying vocabulary before the class; however, during the other three stages, in which some form of vocabulary testing was administered, many more students were studying vocabulary when I entered the classroom. On average, less than 25% of the students were studying before the class began during the no-testing stage, but this increased to between 50% and

60% during the other stages, with the most noticeable increase observed during the first and last testing conditions—student-to-student verbal test not affecting course grade, and student-generated written test affecting course grade. In addition, as discussed below, some students stated in their open survey and interview responses that the test conditions increased the amount of deliberate vocabulary study they did. These findings lend support to the quantitative results showing that the inclusion of the tests significantly increased motivation as measured by amount and type of deliberate vocabulary study carried out. However, to obtain a fuller understanding of the motivational effects of the test conditions, further analyses were carried out on the survey, interview, and observation data.

Although it was clear that the students had a positive view of the student-generated tests and that motivation levels increased as a result of their inclusion in the courses, the extent to which the tests increased motivation and amount of study was not clear from the survey data alone. For example, when answering the open survey question of *What do you think of the weekly spoken vocabulary test system?*, 67 students (92%) gave positive answers, but few of them stated how it affected their study choices. When responding to the same survey question regarding the written test, the results were similar with 61 students (91%) expressing positive onions about the written test, but few explicitly stating that the tests affected their motivation or study time. When examining the answers more closely, it was found that only 14 students (19%) gave answers indicating increases in motivation for deliberate vocabulary study due to the presence of the student-to-student verbal test system, and only six students (9%) stated that they studied more due to the written test.

The data suggest that the tests increase motivation and study time for some students, but the vagueness of the responses to the open-ended survey questions causes problems when assessing the true effect. Table 29 shows example survey responses.

Table 29. Example of Participant Open Responses Indicating Increases in Motivation or Study Time due to Test Conditions

Student-to-student verbal test	Student-generated written test
I had to study more and more	It is good pressure for me
I tend to be lazy, so the test was good for me to study more	It made me memorize words more accurately
I study for the test	I study hard for the test
If we have the test, we will study hard	I thought I have to study hard.
I had high motivation by the test system	It was good opportunity for me to study vocabulary harder
It motivates to study	I think it's good because I also need to study how to write

When given the choice to select any number of nine statements regarding the tests, 41 students (55%) chose *the vocabulary tests make me study more* option. Additional data were collected through informal interviews with 12 students and the analysis of the interview data supported the latter part of the survey data. Of those interviewed, seven students (58%) stated that they studied more frequently under all three testing conditions than in the no test condition. Two students stated that they studied more in the first few weeks after the introduction of the verbal test (condition 2), but then their amount of study time and activities decreased progressively before increasing again when the written test was announced. The remaining three students stated that the tests either affected the amount of deliberate vocabulary study they did very little, or the effort and time they invested was more dependent on how busy they were than whether there were vocabulary tests or not.

In summary, the survey, observation, and interview data indicated that while most participants' have generally positive perceptions of the student-generated tests, the inclusion of the tests increased motivation levels for some students but had little or no effect on others. The survey data indicated that these tests increased motivation and time-on-task for 55% of the participants, a result that supports the statistical data discussed earlier. For example, the results of the Freidman test and analysis of the descriptive statistics showed that while there were significantly higher motivation levels in all three vocabulary testing conditions, some students did not study under any of the conditions, while some did not change their study habits over the course of the study.

Next, the survey, additional observation, and interview data were analyzed and triangulated to identify whether the tests were more intrinsically or extrinsically motiving. The results indicated that the verbal test system was more intrinsically motivating whereas the written test was extrinsically motivating.

First, the survey responses were too vague overall to make a clear distinction with regards to the verbal test system. For example, most of the 67 students who expressed positive opinions about the verbal test used general phrases such as such as *it is great*, and *it is good*; thus, they did not explicitly differentiate between the two types of motivation, but they leaned towards intrinsic motivation. In addition, six participants specifically referred to intrinsically motivating factors such as it is *fun*, *exciting*, and *enjoyable*. However, a few respondents used phrases that indicate extrinsic motivation such as *it is useful,* or *I have to/had to study*, but only one specifically mentioned an extrinsic factor; “I tried hard when score is watched by my friend” (avoiding loss of face). Therefore, because the survey data results are unclear, further analyses were carried out with the interview and in-class observation data. It appeared that the verbal test was more intrinsically motivating than extrinsically motivating. While only six participants specifically referred to intrinsically motivating factors in the survey, during the interviews, many students said that the test was fun, enjoyable, or felt like a kind of game. In addition, the atmosphere observed in class during the verbal testing procedure (every week for a six-week period) was lively and energetic, with many students laughing and smiling throughout. Many students who were not engaged or energetic during other in-class activities were engaged during the verbal test. In addition, the survey and interview data indicated that many students thought that the tests were a practical and enjoyable method of studying, reviewing, or learning more about the target vocabulary items, but they did not view the system as an actual or serious test. For example, except for the one respondent mentioned above, none of the survey or interview respondents stated that they were worried about the test or how it affected their grade.

The written test data show that it is more extrinsically motivating than intrinsically motivating. First, no students mentioned any intrinsically motivating factors when answering the survey questions, and most of the responses relating to the effect of the written test on study habits involved answers with verbs relating to extrinsic motivation such as I *have to study hard*, *it made me memorize words more accurately*, and *it is difficult so I must study more*. In addition, when questioned, all of the interviewees stated that they did not enjoy the tests but thought that they were useful or necessary. In addition, before the tests, in-class observations revealed that many students seemed nervous. When asked about this issue in the interviews, the participants stated that they felt nervous or were worried about their test scores, but this was due to the fact that it was a written test more than the fact that it could affect their grade. Overall, the survey, observation, and informal interview data suggest that the written test is viewed as more of a standard test, and therefore increases time-on-task because of external motivation.

In summary, the observation and interview data indicated that (a) the inclusion of the student-generated tests increases motivation levels for some but not all students, and (b) the verbal tests mainly increase intrinsic motivation levels while the written test is extrinsically motivating.

Reliability of Students as Examiners

Three types of test data were analysed to answer research question 3: student-student verbal test scores, teacher-administered verbal test scores, and student-generated written test scores. Due to time restrictions I could not administer the verbal test to all students involved in the first part of motivation part of the study, and some students were absent for the written test; thus, the participant sample size was 50.

Four types of statistical analysis were carried out; Friedman's test, Kendall's coefficient of concordance (Kendall's *W*), and Spearman's rank-order correlation and intraclass correlation coefficient analysis. The results of the Friedman test analysis showed whether there are significant differences between the student-administered and teacher-administered test scores, and the other two analyses indicated how closely each test type was related. This multi-step analysis enabled the evaluation of how reliable students are as examiners, thus indicating the extent to which the test scores can be trusted and used for assessment purposes.

The descriptive statistics for the Friedman test analysis are shown in Table 30.

Table 30. Descriptive Statistics for the Three Vocabulary Tests

	Student-student verbal test	Teacher-administered verbal test	Student-generated written test
M	14.47	14.5	14.18
Min.	7	8	8
Max.	19	18	18
SD	2.95	2.17	2.25
SE	.29	.26	.23
Mdn	15.00	15.00	14.00
95% CI (M)	[13.63, 15.24]	[13.86, 15.08]	[13.54, 14.82]
95% CI (Med.)	[14.00, 16.00]	[14.00, 15.00]	[14.00, 15.00]

Note. *N* = 50. CI (M) = Confidence interval of the mean. CI (Med) = Confidence interval of the median.

As seen in Table 30, the mean, minimum, maximum and median scores are all very similar. This is actually very surprising as it had been assumed that there would be a bigger difference between the test scores. It is also surprising that the teacher-administered test score has a slightly higher mean than the other two tests as the teacher was initially deemed to be the strictest and most accurate rater. In addition, the student-

student verbal test has the lowest minimum score, something which is also slightly surprising, and while the written test appears to be the most difficult, as expected, the difference is very slight. Also, considering the minimum, maximum and average test scores (between 14.18 and 14.5 out of 20, between 70.9% to 72.5%), it seems that the data are what would be expected given the study count data analyzed for research question 1, where some students did not study at all and others studied a great deal.

After examining the descriptive statistics, a Friedman test was run to investigate whether there were significant differences in the vocabulary scores of the three tests. As determined by the Friedman test, while there were some differences between the various test scores, they were not statistically significant, $\chi^2(2) = 1.926$, $p = .38$. While not necessary, pairwise comparisons were made using the Wilcoxon signed rank tests, and the results confirmed those of the Friedman test as none of the individual tests differed significantly to any of the others ($p > .017$ for all comparisons). The results show that the distributions of the spoken test, written test, and teacher administered test are not significantly different.

Next, Kendall's *W* was run to investigate whether there was agreement between the three tests on the participants vocabulary knowledge ($N = 50$). For all three tests, the vocabulary knowledge of the 50 participants was determined according to a continuous scale of 0 to 20. The three vocabulary tests statistically significantly agreed in their assessments, $W = .852$, $p < .001$. As Kendal's W values of .10 indicate a small agreement, .30 represents moderate agreement, and > .50 indicates strong agreement

(Field, 2018, p. 328; Laerd Statistics, 2019), the results showed strong agreement amongst the three tests.

Next, Spearman's rank-order correlation was run to assess the strength of the relationship between the three vocabulary tests ($N = 50$). Preliminary analyses showed the relationship to be monotonic, as assessed by a visual inspection of a scatterplot. The Spearman's correlation coefficients (r_s) tended to be smaller than those of a Pearson correlation so, for example, a r_s value above .70 indicates a high correlation (Laerd Statistics, 2019). As determined by the Spearman's correlation analysis, there was a statistically significant positive correlation between the three tests. First, there was a statistically significant, strong positive correlation between the student-student verbal test scores and the teacher-administered verbal test scores, $r_s(48) = .77$, $p < .001$. Second, there was a positive significant correlation between the student-student verbal test scores and the student-generated written test scores, $r_s(48) = .77$, $p < .001$. Third, there was a positive significant correlation between the teacher-administered verbal test scores and the student-generated written test scores, $r_s(48) = .81$, $p < .001$.

The final analysis involved calculating the intraclass correlation coefficient to investigate inter-rater (inter-test) reliability. ICC estimates and their 95% confident intervals were calculated using SPSS 26 based on a mean-rating ($k = 3$), absolute-agreement, one-way random-effects model. Values less than .50 indicate poor reliability, values between .50 and .75 indicate moderate reliability, values between .75 and .90 indicate good reliability, and values greater than .90 indicate excellent reliability (Koo & Li, 2016; Laerd Statistics, 2019). As determined by the analysis, a good to excellent degree of reliability was found between the vocabulary measurements; average measure ICC = .920 with a 95% confidence interval from .87 to .95, $F(1, 49) = 12.56$, $p < .001$.

In summary, given the non-significant Friedman test ($p < .017$), the Kendall's W of .85, ($p < .001$), the strong positive correlations ($r_s = .77$, $p < .001$), and the good to excellent degree of reliability (ICC = .92, CI = .87 to .95 , $p < .001$), the student-generated tests are reliable; thus, the students were reliable examiners.

Discussion

Motivational Effects

The first main finding of this experiment is that without vocabulary testing, the participants engaged in little deliberate vocabulary study. On average, the students studied approximately one out of three weeks and completed less than six activities during the same period. Seventeen students (20%) did not study at all and 29 students (34%) studied only one out of the three weeks. Based on activity completion counts and frequency data, 30 students (35%) completed zero or one activity in three weeks, and based on the results of the pilot study and informal interviews, 35% of the participants were not motivated enough to devote more than five to seven minutes to deliberate vocabulary study in three weeks. In addition, 72% of the participants completed only six or fewer activities, which equates to approximately 30 minutes or less of vocabulary study in three weeks. Only 11 students (13%) spent one hour or more and completed 12 or more activities during the three weeks. These results paint a bleak picture, especially given that 98.6% of the 73 students who responded to the survey ranked vocabulary as very or extremely important to their English ability and 59 (79.7%) reported medium to high motivation levels for vocabulary study. These results support previous findings that many learners struggle to find or maintain motivation for deliberate vocabulary study (Al-Murtadha, 2014; Ng & Ng, 2015; Tanaka, 2017; Zheng, 2012). In addition, based on an understanding of the importance of frequency, repetition, and time-on-task (Brown, 2012; Eckerth & Tavakoli, 2012; Schmitt, 2008), the results indicate that most learners did not devote enough time or effort to meaningfully build their vocabulary.

The next main finding is that motivation levels, as measured by study counts and activity completion counts, significantly increased under all test conditions compared to the no-testing condition. The statistically significant differences of the Friedman tests confirmed that the participants studied more times and completed more activities in each testing condition than when they were not tested. This finding supports the position that regular testing increases motivation in educational contexts (Adapa, 2015; Butler, 2010) and more specifically for deliberate vocabulary study (Adapa, 2015; Wilkinson, 2015). This finding is especially significant because no research has been found that explicitly show the effect of (a) vocabulary testing in general or (b) different test conditions on motivation for deliberate vocabulary study. Although Moir and Nation (2002), Ng and Ng (2015), Thompson (2014), and Wilkinson (2015) have published findings indicating that testing can increase motivation for vocabulary study, they did not test this idea experimentally. As Tseng and Schmitt (2008) stated, “only a small number of studies have been undertaken to examine the role of motivation in vocabulary learning … and further research needs to be undertaken to systematically examine its effect on the vocabulary learning process” (pp. 358–359); therefore, this finding fills an important gap in the vocabulary learning literature.

Another finding is that the verbal tests are intrinsically motivating and assigning grades to the tests does not increase extrinsic motivation, but it might reduce intrinsic motivation. The testing condition that had no effect on the participants’ course grade produced the highest motivation, although the differences were not statistically significant. An analysis of the quantitative and qualitative survey, observation, and interview data revealed that this test condition was the most intrinsically motivating. Taking all of the data together, levels of intrinsic motivation increased due to the inclusion of the test system and this positively affected study time and effort; enjoyment had a more positive effect than threat. This finding supports previous work concerning the importance of intrinsic motivation in general (Deci & Ryan, 1985, pp. 13–29; Dörnyei, 2001, pp. 11, 51; Harackiewicz & Sansone, 2000; Harackiewicz & Sansone, 2000, pp. 79–80) and for vocabulary learning (Schmitt, 2008). In addition, based on the survey, observation, and interview data, the students enjoyed the peer-supported or cooperative nature of the verbal test system, supporting previous findings regarding the benefits of such activities (Masson, 2012; Ruegg, 2013).

On the other hand, although there was no statistically significant effect, assigning part of the course grade to the verbal test scores reduced study time somewhat, so it might have decreased intrinsic motivation while not increasing extrinsic motivation. The fact that after assigning grades to the verbal test, study levels did not increase could have been due to the novelty of the test system wearing off, but some students stated that the test should not be used for grading because it can be unfair; some students appeared to select easier words for the test, the test difficulty varies, students run out of time, and some students are more helpful and give too many hints, which is a form of cheating. In addition, no students stated that they worried about the grades on the test or how it would affect their course grade. Therefore, assigning grades to the test does not seem sensible; it does not increase extrinsic motivation but can reduce intrinsic motivation (Deci, 1971; Dörnyei, 2001, pp. 94, 132; Workman & Williams, 1980). This conclusion is logical given that the students generally had positive views of the tests and felt that they are a worthwhile and necessary vocabulary activity, something that might have increased intrinsic motivation levels and made the grading component unnecessary.

However, if teachers feel the need to use the tests for grading purposes, as is often the case when administering tests (Butler, 2010; Fulcher & Davidson, 2007, p. 192), the survey and interview results indicate that the student-generated written test system is most suitable for this. When answering the open-ended questions, many participants stated that they thought that the written test (a) is the most accurate measure of their vocabulary knowledge, (b) most closely resembles what they view a vocabulary test to be, and (c) is the most reliable because they cannot receive assistance from their partners.

Finally, the survey results support the inclusion of student-generated tests as part of the language-focused learning component of language courses (Nation, 2007). In addition to enjoying the verbal test system, many students stated that both forms of student-generated tests (verbal and written) were (a) useful or necessary, (b) helpful for noticing gaps in their vocabulary knowledge, (c) provided a good opportunity to study or review target items, and (d) offered chances to gain broader knowledge of the target items, for example, pronunciation, word family knowledge, and use in sentences. Therefore, given the positive motivation results combined with the importance of noticing (Nation, 2013, p. 103; Schmidt, 1990; Schmitt, 2008), repetition (Folse, 2006, Horst, Cobb, & Meara, 1998; Mizumoto, 2013), engagement (Cobb, 1999; Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2014), and involvement load (Hulstijn; 1992; Hulstijn, Hollander, & Greidanus, 1996; Nagy et al., 1985; Watanabe, 1997), there is a great deal of support for the tests used in this study. In addition, the tests are supported by their focus on building both breadth and depth of vocabulary knowledge (Nation, 2013, pp. 13, 49; Schmitt, 2010, pp. 15–17, Webb, 2020, p. 6).

Reliability of the Student-Generated Vocabulary Tests

The results indicated that the student-generated tests reliably measured vocabulary knowledge, especially if we consider the teachers vocabulary measure to be accurate. For example, if there had been statistically significant differences between the student and teacher administered test scores, or the student-generated scores were not strongly correlated with those of the teacher, this would show the unreliability of students as examiners. However, the results indicated that the students were reliable examiners, both when administering and scoring the verbal tests and when marking the written tests. This result offers support for viewing the peer-to-peer tests as reliable assessment tools, or as a method of providing learners with accurate feedback on their vocabulary development, something students need and find beneficial (Masson, 2012; Metcalfe & Kornell, 2007; Resendes, Chen, Acosta, & Scardamalia, 2013; Ruegg, 2013).

The results of the statistical analyses indicated that teachers can use one or both types of tests for grading purposes, but the survey and interview data indicated that the verbal test should not be used for such purposes. When answering the open-ended questions, many participants stated that the written test (a) is the most accurate measure of their vocabulary knowledge, (b) most closely resembles what they view a vocabulary test to be, and (c) is the most reliable test because they cannot receive hints or help form partners. Therefore, the verbal tests should be used to increase intrinsic motivation and provide opportunities for repeated study and review (Schmitt, 2008), while the written test is more appropriate for grading purposes.

The tests are especially useful in courses in which students self-select their target vocabulary, but teachers want to administer tests for grading purposes. Regular testing is widely recommended (Butler, 2010; Fulcher & Davidson, 2007, pp. 192–193), especially if it does not take too long (Nation, 2013, p. 470), but so is the vocabulary self-selection strategy (Harmon et al., 2008; Mantiri, 2015; Masoudi, 2017; Ruddell & Shearer, 2002). However, these two approaches are in direct conflict with each other because the self-selection strategy makes vocabulary testing difficult. The results of the reliability analysis indicated that the student-generated tests offer a suitable solution to this problem; teachers can allow learners to adopt the self-selection strategy while also having a method of assessing each student’s vocabulary knowledge and development.

CHAPTER 6

DISCUSSION

The results of the experiments discussed in this study add valuable findings to the vocabulary literature, especially in terms of the efficacy of word cards, and effects of testing on motivation for deliberate vocabulary study. While some findings of the first two experiments are supportive of previous findings regarding word cards (Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014) and deliberate learning techniques in general (Cobb, 1999; Laufer & Rozovski-Roitblat, 2014; Nakata, 2008; Waring, 2004), they also call into question some research-based conclusions previously made (Dodigovic, 2013). In addition, the results of Experiment 3 support the view that regular testing can increase motivation for vocabulary learning (Al-Murtadha, 2014; Schmitt, 2008; Zheng, 2012), but the nature of the tests and measure of motivation involved in this study make the findings quite different to previous research. Most importantly, the main results of the three experiments add new and valuable information to the vocabulary acquisition literature.

The first main finding of this study came from Experiment 1 and is that the process of making word cards results in statistically and pedagogically significant learning outcomes in both the short-term and long-term, a finding that adds to the research-based support for word cards (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Nikoopour & Kazemi, 2014). However, as no previous studies were found that isolated the effects of the word card making process, this is the first study offering empirical support for the statement “We know from experience that after making the cards, learners can recall from 30-70% of the meanings on the first run through the cards” (Nation & Webb, 2011, p. 42). The results of Experiment 1 were that immediately after making 15 word cards in 15 minutes, learners were able to recall up to 70% of the target items. In addition, with no further study being carried out, participants were still able recall up to 50% of the target items two weeks later, thus, the results support the views held by researchers such as Mondria and Mondria-de Vries (1994), Nation (2006), and Webb (2005) that by writing the cards, learners pay considerable attention to form and meaning and spend more time than they might have if they were given premade materials, both of which contribute to the development of longer-term retention of the new vocabulary knowledge. However, as significant attrition did occur between immediate and delayed posttests, the findings support the assertion that repeated study is vital if vocabulary knowledge is to be committed to longer-term memory (Nakata, 2020, pp. 307–311; Nation, 2013, p. 102; Schmitt, 2008) and are similar to previous research findings that reported significant vocabulary knowledge attrition when no further deliberate vocabulary study was carried out for a few weeks (Kuo & Ho, 2012; Oberg, 2011; Nakata, 2008; Waring, 1997, 2004). In addition, given the simple nature and low task-induced involvement load of the controlled word card making task involved in Experiment 1, the results lend support to the view that tasks requiring only shallow processing are better at helping learners commit vocabulary knowledge to short-term memory than developing long-term memory traces (Craik & Lockhart, 1972; Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998). Given the above, in answer to the first two research questions regarding the short-term learning outcomes of making word cards and subsequent levels of attrition if no further study is carried out, the finding is that the process of making word cards results in impressive gains in short-term knowledge of target vocabulary but much of this knowledge is lost in a relatively short period of time if no further study is carried out.

Another finding from Experiment 1 is that students view making word cards as a simple, quick, and effective deliberate study technique, which is similar to findings by Hung (2015), Kuo and Ho (2012), and Nikoopour and Kazemi (2014). The participants in this study viewed the process of making the cards as simple and efficient, but they also stated that it required enough time, effort, and attention to detail to commit the new information to memory (Mondria & Mondria-de Vries, 1994). In addition, the students liked the two-sided nature of the cards, as aspect that adds affective support for one of the often-reported benefits of word card design (Mondria & Mondria-de Vries, 1994; Nakata, 2008; Nation, 2013, p. 451; Schmitt, 2008). However, none of the interviewees stated that they enjoyed making word cards, a finding that provided more evidence to the commonly held belief that deliberate vocabulary techniques are not generally intrinsically motivating (Al-Murtadha, 2014; Elley, 1989; Wilkinson, 2017).

Overall, the major findings for the first experiment are that (a) students have a positive view of the process of making word cards, and (b) the process results in impressive gains in short-term knowledge of target vocabulary, but much of this knowledge is lost in a relatively short period of time if no further study is carried out.

However, this was a very controlled, somewhat unnatural, and cross-sectional word card study, so the results of Experiment 2 must be taken into consideration.

For example, the results of Experiment 2 call into question the benefit of making word cards as one of the main findings is that the process of making word card itself does not seem to offer any benefit over studying from premade word cards, at least if learners are not self-selecting their target items. In fact, when given no choice in target item selection (self-made no-choice condition and premade condition), the premade word card treatment outperformed the self-made treatment, especially in the short-term. This finding supports those of Dodigovic (2013), and Barcroft (2006, 2007b) regarding the inefficacy of copying and writing tasks. In addition, the findings support the statement made by Folse (2006) that "the depth of processing needed to create one's own cards might not be as conducive to retention as is the number of word retrievals required" (p. 273). The results of the immediate posttest showed significantly better learning outcomes for the premade card treatment on both the active and passive recall measures ($p < .001$), with a difference of 2.5 and 1,2 words out of 15 respectively. On average, immediately after the initial treatment stage, 40% of the words encountered when making the word cards were actively recalled, but it was 60% for the words encountered in the premade treatment. Also, 57% of the words in the self-made treatment were passively recalled, but 65% were passively recalled under the premade word card treatment. These immediate posttest results confirm the efficacy of making word cards, but also indicate that if time-on-task is controlled, receptive tasks that offer more repetitions (chances for retrieval) are more effective than time consuming productive tasks (Barcroft, 2004, 2006, 2007b; Webb;

2005, 2018). On the other hand, there was little difference between the two methods at the delayed posttest stage, that is, after eight weeks studying from the cards under natural conditions and then not seeing the words for two weeks. The difference on the delayed posttest active recall measure was only 0.5 words out of 15 ($p > .05$), and that of the passive recall measure was 1.35 words ($p < .001$).

Overall, while the differences do not seem so significant from a pedagogical viewpoint and are only statistically significant for the passive recall delayed posttest measure, the premade method produced better learning outcomes. This indicates that the process of making word cards is not so important under more natural and longitudinal conditions where students have the opportunity to make strategic decisions about studying from word cards. For example, the observation and interview data revealed that, as also noted in the literature, some students study more than others (Webb, 2005), and some students use a broader range of strategies (Kojic-Sabo & Lightbown, 1999; Mizumoto, 2010; Moir & Nation, 2002; Wilkinson, 2017) such as shuffling their packs more often, carrying out an expanded spaced-retrieval strategy, and choosing to combine receptive and productive recall practice, and these choices seem to outweigh the method by which the word cards were created. In addition, the results of the survey and interview data analyses indicate that the statistical learning outcome results might have been affected by motivational factors relating to students' preferences. When asked which style of word cards they preferred, 69% of participants (51 respondents, $N = 74$) stated that they preferred studying from premade cards such as the ones used in this study or those available on smartphones, a finding that supports those of Dodigovic (2013) and is similar to those of Reynolds and Shih (2019). The main reason given was that it is easier and takes less time, but some students also stated that it gives them more time to focus on studying the words. Example comments concerning the participants' perceptions of premade word cards included "I don't know how to make good words cards", "It is hard for me to find unknown words. I also do not have much time to do that," "It's troublesome for me that making card by myself," and "Student can spend time for the vocabulary practice instead of making vocabulary cards." More students preferred the premade approach than the self-made one, and this could have resulted in the better learning outcomes (Dörnyei, 2001, pp. 2, 5; Ng & Ng, 2015). The results discussed above provide fairly strong support for premade cards over self-made, but given that the two conditions were very controlled and unnatural word card methods, the results must be considered in relation the learning outcomes of the third treatment, the self-selection and self-made word card treatment, which was the most natural word card method.

When comparing all three word card treatments, the results were that the self-selection and self-made word card treatment resulted in significantly higher long-term active and passive recall learning outcomes than the other two methods, a finding that calls the superiority of premade cards into question. The results of the delayed posttest analysis were that the self-selection and self-made word card treatment resulted in between 14.7% and 18.4% more active recall knowledge and 11% and 20% more passive recall ability than the other two methods. This shows that when learners are given autonomy in target item selection, and then make their own word cards, the learning outcomes are considerably better, supporting the views that autonomy and self-selection

(Brown, 2012; Cobb, 1999; Laufer, 2014; Mizumoto, 2013), and levels of processing and involvement load (Eckerth & Tavakoli, 2012; Keating, 2008; Laufer & Rozovski-Roitblat, 2014; Newton, 2013) are all important predictors of vocabulary learning.

Next, regardless of the differences in learning outcomes attributable to the three word card conditions, the main finding of Experiment 2 is that all the word card methods resulted in significant vocabulary knowledge gains in both the short- and term-term, supporting previous research (Elgort, 2011, Nakata, 2008; Kuo & Ho, 2012; Reynolds & Shih, 2019). In terms of the short-term outcomes, after only 15 minutes of making word cards or studying from premade cards, the students demonstrated statistically significant and pedagogically impressive gains on both the immediate active and passive recall posttest measures. Immediately after the initial treatment stage, mean active recall vocabulary scores ranged from six to nine words out of 15 (40-60%) and the participants were able to passively recall between 8 to 10 words (57% and 65% of the target items), respectively. In addition, the delayed posttest results showed that all three word card methods were effective for developing longer-term memory of the target items. After studying the words for eight weeks under the three conditions, and then not seeing the words for two weeks, participants could still actively recall between 49% and 64% of the target items, and passively recall between 53% and 73%. Thus, these results add to the growing body of empirical evidence showing that word cards are an effective deliberate study technique (Dodigovic, 2013; Komachali & Khodareza, 2012; Kuo & Ho, 2012; Reynolds & Shih, 2019). In addition, as all word card conditions offered students the chance for multiple encounters with target items, and the opportunity for both active and passive recall or retrieval, the findings concur with the literature citing the importance of each process (Heidari-Shahreza & Tavakoli, 2012; Horst, Cobb, & Meara, 1998; Mondria & Mondria-de Vries, 1994, Nation, 2013, pp. 107–110).

Furthermore, in terms of students' perceptions, the results of survey and interview data analysis in Experiment 2 supported those of Experiment 1, and solidified the finding that students have a positive view towards using word cards for deliberate vocabulary study. These findings are supportive of those found in some previous studies such as Hung (2015), Komachali and Khodareza, (2012), Kuo and Ho (2012), Nikoopour and Kazemi (2014), but are contrary to those found by Dodigovic (2013), Emrah and Omur (2010), and Reynolds and Shih (2019). The data analysis revealed that 89.2% of the participants in Experiment 2 (N = 74) rated word cards as effective. When providing further details through an open-ended survey question, many participants indicated that word cards are quick and easy to make and use. In addition, many participants commented on the convenience of word cards, noting that they are easy to carry and can be studied anywhere, a feature that has been noted in the literature (Emrah & Omur, 2010; Nikoopour & Kazemi, 2014; Schmitt, 2008). Moreover, when interviewed, some students mentioned that they prefer word cards over vocabulary notebooks or textbooks such as those they used in junior or senior high school English courses, findings similar to those of Hirschel and Fritz (2013) and Nakata (2008). Also, the interviewees understood that the two-sided nature of word cards allows for more active and passive recall than a vocabulary notebook where the L1 and L2 forms are placed side-by-side. However, because only one participant commented on the intrinsic motivation factor, "motivate up studying vocabulary… It is interesting," the students clearly did not feel that word cards are an enjoyable or intrinsically motivating activity, supporting claims made in previous research (Emrah & Omur, 2010; Kojic-Sabo & Lightbown, 1999; Reynolds & Shih, 2019; Wilkinson, 2017).

Another finding of the second experiment is that the Involvement Load Hypothesis could be a good predictor of the vocabulary learning outcomes of deliberate study methods, not just incidental learning tasks. However, the results indicate that this might only be the case if the tasks being assessed differ substantially in terms of their task-induced involvement loads and time-on-task is not controlled. For example, given the fact that the premade word card treatment, which had the lowest task-induced involvement load, performed significantly better than the self-made no-choice treatment, especially in the short-term where time on task was controlled, the ILH was not an accurate predictor of the learning outcomes of these two word card methods. However, the two treatments differed only slightly in their task-induced involvement loads (2 and 1 respectively), so it could be that (a) they did not differ enough to make accurate predictions and (b) the individual learner choices and strategies played a more important role, something to be expected given the importance of such choices and strategies (Barcroft, 2009; Kojic-Sabo & Lightbown, 1999; Mizumoto, 2010; Moir & Nation, 2002). On the other hand, the self-made with choice treatment, which had a considerably higher task-induced involvement load (task-induced involvement load = 5), produced the best long-term learning outcomes, where time-on-task was not controlled. The positive differences between the student-selected with choice treatment (task-induced involvement load = 5), and the two controlled word card methods (task-induced involvement load = 1 and 2) at the delayed posttest stage were significant for both dependent vocabulary measures. Given that the student-made with choice treatment had the highest task-induced involvement load (Laufer & Hulstijn, 2001), and the delayed posttest was the most accurate measure of learning outcomes (Nation & Webb, 2011, pp. 233, 278), the results lend further support to the ILH (Keating, 2008; Laufer & Hulstijn, 2001; Laufer & Rozovski-Roitblat, 2014; Pichette et al., 2011), but more specifically suggest it could be a good predictor of the vocabulary learning outcomes of deliberate study methods, but only if their involvement load differs substantially and time-on-task is not controlled. This is a finding similar to that of Keating (2008), although his research focus was different, and his study compared both incidental and deliberate learning tasks.

Another important finding that was present in both Experiments 1 and 2, is that word card study results in better passive recall ability than active recall. Not only were the passive recall mean scores higher on both the immediate and delayed posttests in Experiments 1 and 2, but the passive recall knowledge was less sensitive to attrition over time. These findings support the widely held view that it is easier to develop passive knowledge than active (Folse, 2006; Webb, 2005, 2009) are similar to findings of previous deliberate vocabulary learning studies (Komachali & Khodareza, 2012; Kuo & Ho, 2012; Waring, 1997; Webb, 2009). These results confirm the assertion that language teachers must include activities and tasks that focus on building productive vocabulary knowledge (Webb, 2005, 2009).

In summary, the main findings of the word card studies carried out as part of this study are that (a) word cards are viewed positively by learners, (b) word cards are an effective deliberate learning technique but are better at developing passive recall knowledge than active, (c) the process of making word cards is not so important or beneficial when learners have no autonomy in target item selection, but the task-induced involvement load is an important factor. In addition, the findings indicate that the involvement load hypothesis can be a good predictor of the learning outcomes associated with deliberate study techniques if they differ substantially in their task-induced involvement load and time-on-task is not controlled. However, as discussed in the limitations section, this finding should be read with caution and further research is needed into this area.

Next, the main findings of Experiment 3 concerning the effects of vocabulary testing on motivation for deliberate vocabulary study are discussed. While most of the findings are unrelated to Experiments 1 and 2, some of the results of the data analysis affect the conclusions drawn above regarding premade word cards.

The first main finding of Experiment 3 is that without any form of vocabulary testing, participants engaged in little deliberate vocabulary study, thus demonstrating low levels of either intrinsic or extrinsic motivation for deliberate vocabulary learning. On average, the students studied approximately only one out of three weeks and completed less than six activities during the same period. 20% of the participants did not study at all and 34% studied only one out of the three weeks, indicating either no or low levels of motivation (see Table 25). Based on activity completion counts and frequency data, 35%

completed no or only one activity in three weeks, meaning they spent no more than five to seven minutes deliberately studying vocabulary over a three-week period. In addition, 72% of the participants completed only six or fewer activities, which equates to approximately 30 minutes or less of vocabulary study in three weeks.

These results show a clear lack of intrinsic or extrinsic motivation for deliberate vocabulary study despite 98.6% of the participants who responded to the survey stating that vocabulary is important to their English ability and 79.7% reporting to have medium to high motivation levels for vocabulary study. This finding supports previous research indicating that many learners struggle to find or maintain motivation for deliberate vocabulary study (Al-Murtadha, 2014; Ng & Ng, 2015; Tanaka, 2017; Zheng, 2012). In addition, based on an understanding of the importance of frequency, repetition, and time-on-task (Brown, 2012; Eckerth & Tavakoli, 2012; Watanabe, 1997; Schmitt, 2008), the results indicate that most learners did not devote enough time or effort to meaningfully build their vocabulary.

This finding also calls into question the pedagogical support for using premade word cards, at least in courses with low motivated students that do not include regular testing. The results of Experiment 2 indicate that premade cards could be superior to self-made ones, a claim that was made by Dodigovic (2013), but the low motivation levels found in Experiment 3 indicate that this needs serious consideration. For example, if students are not tested on target vocabulary, there is little way to be sure they have autonomously studied from premade materials, and the results of Experiment 3 indicate that many will not do so, thus increasing the likelihood of limited or no learning outcomes for at least some students in a course. On the other hand, the inclusion of self-made word cards is likely to result in at least some learning outcomes for all students. This is because having learners make their own cards guarantees at least some level of engagement and study, and as shown in Experiment 1, this will result in some positive learning outcomes. Therefore, the question of whether to use self-made or premade deliberate study materials is reliant on the motivation levels of the students, and hinges on the teacher's ability to accurately assess motivation, something that is not always possible based on self-report data as this study shows.

The next main finding is that all test conditions significantly increased motivation as measured by study counts and activity completion counts. The statistically significant differences of the Friedman tests confirmed that the participants studied more times and completed more activities in each testing condition than when they were not tested. This finding supports the position that regular testing increases motivation in educational contexts (Adapa, 2015; Butler, 2010; Fulcher, & Davidson, 2007; pp. 222–224) and specifically for deliberate vocabulary study (Adapa, 2015; Wilkinson, 2015). This finding is especially important because no research has been found that examined the motivational effects of foreign language vocabulary testing in general, nor were any found that examined the motivational effects of different test conditions. Although Moir and Nation (2002), Ng and Ng (2015), Thompson (2014), and Wilkinson (2015) have concluded that testing can increase motivation for vocabulary study, they did not test this idea experimentally. As Tseng and Schmitt (2008) stated, "only a small number of studies have been undertaken to examine the role of motivation in vocabulary learning …

and further research needs to be undertaken to systematically examine its effect on the vocabulary learning process" (pp. 358–359). Therefore, the finding that regular testing increases motivation for deliberate vocabulary learning, even with low motivated learners such as those in this study, fills an important gap in the vocabulary learning literature.

A surprising finding is that the verbal student-to-student test system is intrinsically motivating but assigning grades to the tests does not increase extrinsic motivation. The student-to-student test condition that had no effect on the participants' course grade produced the highest motivation levels, although the differences between the other two grade-weighted test conditions were not statistically significant. An analysis of the survey, observation, and interview data revealed that the verbal student-to-student test system with no grade attached was the most intrinsically motivating test condition. Taking all of the data together, levels of intrinsic motivation increased due to the inclusion of the test system and this positively affected study time and effort; enjoyment had a more positive effect than threat, supporting previous conclusions regarding the importance of intrinsic motivation in language learning contexts (Dörnyei, 2001, pp. 11, 51; Harackiewicz & Sansone, 2000, pp. 79–80) and for vocabulary learning (Nation, 2013, p. 102; Schmitt, 2008). In addition, based on the survey, observation, and interview data, the students enjoyed the peer-supported or cooperative nature of the verbal test system, supporting previous findings regarding the benefits of such activities (Al-Murtadha, 2013; Hung, 2015; Masson, 2012; Ruegg, 2013).

On the other hand, although there was no statistically significant difference, assigning part of the course grade to the verbal test scores reduced the amount of deliberate vocabulary study, indicating that it decreased motivation levels. In this case, based on the survey and interview data, it seemed that assigning grades to an enjoyable and intrinsically motivating activity decreased intrinsic motivation while not increasing extrinsic motivation. For example, some students stated that the test should not be used for grading and no students stated that they worried about the grades on the test or how it would affect their course grade. Therefore, assigning grades to the verbal test system does not seem sensible because creating and maintaining intrinsic motivation is a key determinant of success in foreign language learning (Deci, 1971; Dörnyei, 2001, pp. 11, 51), and if it does not increase extrinsic motivation but reduces intrinsic motivation, overall motivation levels will go down (Workman & Williams, 1980). However, if teachers do feel the need to use the tests for grading purposes, as is often the case when administering tests (Butler, 2010; Fulcher & Davidson, 2007, pp. 192–193), the survey and interview results indicate that the student-generated written test system is most suitable for this. When answering the open-ended questions, many participants stated that they thought that the written test (a) is the most accurate measure of their vocabulary knowledge, (b) most closely resembles what they view a vocabulary test to be, and (c) is the most reliable because they cannot receive assistance from their partners.

Finally, the survey results support the inclusion of student-generated tests as part of the language-focused learning component of language courses (Nation, 2007). In addition to enjoying the verbal test system, many students stated that both forms of student-generated tests (verbal and written) were (a) useful or necessary, (b) helpful for noticing gaps in their vocabulary knowledge, (c) provided a good opportunity to study or review target items, and (d) offered chances to gain broader knowledge of the target items, for example, pronunciation, word family knowledge, and use in sentences. Therefore, given the positive motivation results combined with the importance of noticing gaps in knowledge (Nation, 2013, p. 103; Schmidt, 1990; Schmitt, 2008.), repetition and spaced rehearsal (Horst, Cobb & Meara, 1998; Mizumoto, 2013; Nation, 2013, pp. 107, 451; Schmitt, 2008), engagement (Cobb, 1999; Eckerth & Tavakoli, 2012; Folse, 2006; Laufer & Rozovski-Roitblat, 2014), and involvement load (Keating, 2008; Hulstijn; 1992; Hulstijn, Hollander, & Greidanus, 1996; Nagy et al., 1985), there is a great deal of support for the tests used in this study. In addition, the tests are supported by their focus on building both breadth and depth of vocabulary knowledge (Nation, 2013, pp. 13, 49; Schmitt, 2008; Schmitt, 2010, pp. 224–225).

The final part of Experiment 3 investigated the reliability of the student-generated vocabulary tests used in this study. As mentioned in Chapter 5, in order to investigate how reliable students are as examiners when administered and scoring the verbal and written peer-generated tests, the student-student verbal test scores, teacher-administered verbal test scores, and student-generated written test scores were compared using four types of statistical analyses. By running the four types of analyses, it is possible to evaluate the reliability of the three tests and to assess the reliability of students as examiners, thus determining if the student-generated tests can be used in place of more standardized teacher-generated written tests; something which could prove extremely useful in courses where students are self-selecting their target items. Somewhat surprisingly, the main finding is that when administering and scoring peer-generated vocabulary tests, the participants were reliable examiners.

The results of the statistical analyses indicated that if the teachers' vocabulary measure is accurate, the students were reliable examiners when administering and scoring both the verbal and written tests. Therefore, the peer-to-peer tests represented a reliable vocabulary assessment tool and if they so wish, teachers can use one or both types of tests for grading purposes. This finding is important, but as this is the first study to investigate the matter, it does not clearly connect to any previous research.

Another finding was, based on the analysis of the survey and interview data, was that participants felt the test were an effective method of providing them with accurate feedback on their vocabulary development, something students need and find beneficial (Masson, 2012; Resendes et al., 2013; Ruegg, 2013). When answering the open-ended questions, many participants stated that the verbal test system effectively helped them review their target vocabulary, provided information about vocabulary development, and helped them identify which words needed more attention. However, many participants stated that the written test is the most accurate and reliable measure of their vocabulary knowledge because they cannot receive hints or help form partners. Therefore, while both tests were reliable, the verbal tests should be used to increase intrinsic motivation and provide opportunities for repeated study and review (Schmitt, 2008), while the written test is more appropriate for grading purposes.

The tests are especially useful in courses in which students self-select their target vocabulary, but teachers want to administer tests for grading purposes. Regular testing is widely recommended in various learning contexts (Butler, 2010; Fulcher & Davidson, 2007, pp. 192, 222–224) and in vocabulary learning contexts (Kanayama & Kasahara, 2018). However, so is the vocabulary self-selection strategy (Harmon et al., 2008; Mantiri, 2015; Masoudi, 2017; Ruddell & Shearer, 2002), but these two approaches are in direct conflict with each other because the self-selection strategy makes vocabulary testing difficult. The results of the reliability analysis indicated that the student-generated tests offer a suitable solution to this problem; teachers can allow learners to adopt the self-selection strategy while also having a method of assessing each student's vocabulary knowledge and development.

In summary, the main findings of Experiment 3 are that the student-generated tests (a) increase motivation for deliberate vocabulary learning, (b) offer a reliable method of assessing learners' individual vocabulary development, and (c) are viewed positively by learners. The tests therefore offer a reliable and motivating method for teachers to assess learners' individual vocabulary development and give learners valuable feedback, even if adopting the self-selection strategy.

CHAPTER 7

CONCLUSION

Summary of the Findings

The findings obtained from the three experiments described in this study have led to a number of conclusions regarding (a) the efficacy of different word card methodologies, (b) the effects of vocabulary testing on motivation for deliberate vocabulary study, and (c) students' ability to reliably assess their peers' vocabulary knowledge. Taken together, the results can be used to guide further research and help teachers make informed decisions regarding how to approach the vocabulary component of the language-focused learning strand in foreign language courses.

First, when examining how much initial learning occurs from making word cards and assessing how sensitive this knowledge is to attrition, the findings demonstrated that the process of making word cards results in impressive gains in short- and long-term knowledge of target vocabulary. However, a significant amount of short-term knowledge is lost in a relatively short period if no further study is carried out. Given that the learners were able to recall up to 70% of the target items immediately after making the cards, and up to 50% of this new target item knowledge was retained after two weeks without further exposure to the words, students can acquire substantial knowledge of target vocabulary through the process of making word cards. Therefore, the process of making word cards can be viewed as an effective initial stage of the deliberate study process, but as the results of the second experiment revealed, it must be followed up with further study. While the specific factors responsible for learning outcomes associated with the word card making process were not examined, the process involves sufficient levels of at least some of the following; involvement, engagement, noticing, receptive and productive focus on form and meaning, and time-on-task. However, given the attrition levels between the immediate and delayed posttests, the process of making word cards is not complex or engaging enough to transfer all knowledge processed in short-term memory into long-term memory, so learners must follow up the process with further repeated deliberate study.

Next, when investigating students' perceptions of making word cards, their responses highlight a positive attitude towards the process. Most of the learners viewed the word card making process as quick, simple, and efficient, but they also believed that it requires enough time, effort, and attention to detail to commit the new information to memory. However, although students see the benefits of making word cards, they do not enjoy it; thus, it is not an intrinsically motivating process.

In the second experiment, it was found that when time-on-task is controlled, receptively studying from premade cards is more effective for developing short-term vocabulary knowledge. This finding leads to the conclusion that, at least in the short-term, simple repetitive tasks that offer more chances for retrieval are more effective than time consuming but simple productive tasks. However, the findings for Experiments 1 to 3 together indicated that the use of premade cards is only advisable in situations in which teachers are confident that their learners are highly motivated to engage in deliberate vocabulary study. If students are not intrinsically or extrinsically motivated to carry out deliberate vocabulary study, using premade cards increases the chance that some learners will not study the vocabulary at all, as found in the second and third experiments. On the other hand, having learners make their own cards guarantees a degree of effort, engagement, and time-on-task, and based on the results of Experiments 1 and 2, this therefore guarantees positive vocabulary learning outcomes for all. Ultimately, it is essential for teachers to assess motivation levels and consult with their students before deciding what type of word cards should be used, but if this is not possible, the self-selected and self-made method is recommended.

In addition to the above findings, support for having learners make their own word cards after self-selecting the target items comes from the fact that the self-selected and self-made word card method resulted in significantly better long-term vocabulary learning outcomes. The increased task-induced involvement load gained from having learners self-select target items (i.e., decide they *need* these items), *search* for their L1 meaning, and then *evaluate* the choices and information provided in the dictionaries during the search component significantly increases learning outcomes. As the self-selected and self-made word card method had the highest task-induced involvement load, it was found that the Involvement Load Hypothesis can be used to predict the efficacy of deliberate study techniques, not just incidental tasks. Given these findings, it is recommended that word cards and other deliberate study methodologies be designed or manipulated in such a way as to increase the task-induced involvement load.

Furthermore, the results of Experiment 2 confirmed that many students have a positive view of using word cards for deliberate vocabulary study. Many participants indicated that word cards are quick and easy to make and use, are easy to carry, can be studied anywhere, and are preferred over vocabulary notebooks or textbooks. The affordance for both active and passive recall due to the two-sided nature of word cards was a particularly popular feature. However, many participants expressed a preference for premade word cards, with some of them correctly pointing out that using such cards gives them more time to focus on studying the target items, but most preferred this method because it required less effort. In addition, although students might understand and appreciate the main benefits afforded by making or studying from word cards, this study provides further evidence that, like many other deliberate vocabulary learning techniques, using words cards is not intrinsically motivating.

In addition, while not an original point of focus in the first two experiments, it was found that word cards are significantly more effective at building passive or receptive vocabulary knowledge than active knowledge or knowledge of how to use words in context, even when sentences are added to the cards. Therefore, regardless of the degree to which language teachers support the use of word cards, they must include other tasks that focus on building productive vocabulary knowledge.

The results of the analyses carried out in Experiment 3 paint a clear picture of the positive effects that testing can have on motivation for deliberate vocabulary learning, even with low-motivated learners. After analyzing the data, it became evident that many of the participants had low motivation for deliberate vocabulary study, even though they knew how important vocabulary was to their language development and they had stated that they that were highly motivated to develop their vocabulary knowledge. The low levels of motivation became evident at the beginning of the study because when they were not being tested on the vocabulary that they should be studying, many students did not carry out any deliberate vocabulary study, most studied very little, and only a few studied enough to make meaningful improvements. However, all test conditions in this study significantly increased student motivation, with the analyses confirming that the participants studied more often and completed more activities in each testing condition than when they were not tested. Therefore, the results confirmed that regular testing can be used to increase motivation for deliberate vocabulary learning, and the tests investigated in this study are especially useful in situations in which teachers adopt the vocabulary self-selection strategy with low-motivated learners.

The analysis of the data regarding students' perceptions of the student-to-student verbal vocabulary test system showed that such tests can be intrinsically motivating. This result was primarily due to the cooperative learning nature of the test and because the students enjoyed and valued the feedback they received from their peers by taking the tests. However, if using the verbal test for grading purposes, not only do extrinsic motivation levels remain largely unchanged, the test becomes less fun and enjoyable, and as a result is less intrinsically motiving. Therefore, the verbal test should not be used for grading purposes, but rather as an enjoyable vocabulary review or study activity that provides a valuable opportunity for learners to receive feedback and notice gaps in their vocabulary knowledge. Although the written peer-generated test was viewed as a more standardized, serious, and accurate test, it did not increase extrinsic motivation to an extent that study levels surpassed those under the verbal test condition; students studied more when they were intrinsically motivated by the verbal test system.

Finally, when assessing the reliability of students as examiners or raters of their peers' vocabulary knowledge, the participants were just as effective as the teacher. As a result, if learners self-select their target vocabulary but teachers require a method of assessing and grading students on their vocabulary development, the tests investigated in this study can be used to meet this need. Under normal circumstances in which the vocabulary self-selection strategy is adopted, it is almost impossible for the teacher to test students because each student has their own individual set of target items, but the peer-generated tests offer a solution to this conundrum.

Limitations

There are five major limitations of the experiments carried out in this study that need to be discussed.

First, all of the studies contained a relatively small number of participants who were drawn from only one or two universities; the participants in Experiments 1 and 2 came from two universities, but those in Experiment 3 came from only one. Therefore, the participants were of similar age and had similar educational and socio-economic backgrounds, so the results might not apply to younger or older learners or learners with different cultural or educational backgrounds.

Second, due to time and logistical constraints, only a small sample of participants were interviewed, so while the qualitative data provide interesting insights, the findings might not hold true for all of the participants. Furthermore, due to my lack of Japanese ability, the surveys and interviews were administered in English, and although some participants chose to answer in Japanese, it is likely that more detailed and richer data could have been collected had the surveys and interviews been conducted in Japanese.

Third and possibly the most important limitation of Experiment 1 was that the target items, translations, and example sentences were carefully selected to reduce confounding variables such as word length and part of speech; however, this led to a somewhat unnatural word card making condition. Therefore, while the process of controlling the word card information presented and entered onto the cards increased internal validity, it reduced ecological validity because the self-made no-choice word card making treatment did not accurately represent a real-world card making process. In my experience, while learners are often instructed what types of information to include on the cards, for example, the L2 word, their L1 meaning, an example sentence, learners often self-select the L2 target items and chose what they view to be the most suitable meaning and example sentence. This self-selection process provides learners with more autonomy and engagement and so is likely to be more motivating than the self-made no-choice word card treatment used in Experiment 1. In addition, the self-made no-choice word card making treatment reduced the level of processing involved in making the word cards because the treatment did not allow the learners to search for the L1 meaning of the target words to the same extent as what would occur under normal learning conditions. Nor could the learners choose what they felt was the best L1 meaning or write original sentences; therefore, less semantic processing took place, so learning outcomes might have been reduced. However, although the learners did not have freedom concerning vocabulary selection, they were closely following the usual physical process, such as looking at the unknown L2 item, referring to a dictionary to find their L1 meaning and an example sentence, and entering the information onto the two sides of a card while repeatedly referring to the dictionary to ensure correct spelling and accurate sentences. Therefore, the results show what kind of learning outcomes can be expected from the process of making word cards.

Next, the research designs used in the studies made it impossible to isolate all possible confounding variables. For example, in the word card studies, the individual effects of the receptive and productive process involved, the number of repetitions or retrievals, and the levels of engagement are unknown. Furthermore, the longitudinal design of Experiment 2 increased the chance of other confounding variables affecting the results. For example, because mid-frequency words were chosen, it is unlikely that the learners met those words outside of the experiment, but this could not be guaranteed. In addition, while the participants were asked to select words for the self-selected and self-made treatment from the same frequency level as those used in the other two treatments, no controls could be put in place to ensure this.

Finally, in Experiment 3, it would have been ideal to accurately record time-on-task because this is arguably the most reliable measure of motivation, but this was impossible with the software and resources available, so the best alternative was the study count and activity completion count data. Nevertheless, the study and activity counts are more reliable measures of motivation than self-report data, and they offer a somewhat unique measure of motivation that could prove useful to future researchers. Another potential problem is that some students might have engaged in additional study not using the Quizlet application, and if so, the study count data would be inaccurate. However, based on classroom observations and interviews, this does not seem to be the case because only one student was observed using paper word cards, and her data were not included in the study. Another limitation of Experiment 3 is that due to time constraints during the courses, I was unable to test all students using the verbal test system, thus further reducing the sample size for the reliability analysis.

In summary, while every effort was made to ensure that the research designs led to accurate measurements of the outcomes associated with the variables under investigation, the sample size was small, and it is possible that other variables affected the results. However, the qualitative data were partly used to identify such issues.

Suggestions for Future Research

Future researchers can examine whether these results would generalize to other contexts, for example, with younger learners or learners form different educational or cultural backgrounds. If similar results are found by such replication studies, it would strengthen the findings of both, and make the conclusions drawn here more generalizable to other foreign language learners and educational contexts. However, if different results are found, it would add to the literature citing the importance of individual learner differences to language learning.

In addition, it would be especially interesting to see if the same results are found with more highly motivated learners. It would also be informative if future researchers could include motivation levels as a dependent variable. For example, with regards to the word card learning outcomes, would the results be the same with low motivated, moderately motivated, and highly motivated learners, or is the efficacy of each word card method highly related to motivational levels? To this end, the word card studies described in the first two experiments could be replicated with the addition of a measure of motivation at the beginning of the study and treat this as a categorical variable.

Alternatively, future researchers could compare one or more of the word card methods investigated in this study with alternative word card designs, for example, by comparing premade paper word cards with premade digital ones. Such studies would shed further light on the efficacy of such methods and provide language educators with more empirical support for alternative deliberate vocabulary learning techniques.

Another interesting direction that future word card research could take would be to investigate if repetition is more effective than depth of processing. For example, while treating number of repetitions as an independent variable, researchers could compare simple premade word cards could be with those that include more self-selected information and original sentences.

For the testing and motivation study, it would also be interesting to see if the tests have a similar effect on learners who begin with differing levels of motivation. Therefore, future research investigating the motivational effects of regular vocabulary testing could measure and treat motivation levels at the start of the study as a categorical variable. This type of research could indicate if there are different motivational effects of regular testing learners that have low, medium, and high levels of motivation.

Regarding the reliability students as examiners, and therefore the reliability of the peer-generated test system, researchers could compare these types of test scores with those of standardized or professionally produced vocabulary tests. In addition, research investigating learners' ability to judge their peers' original sentences as part of the peer-to-peer testing would be interesting and provide more insights into the extent to which students can act as test administrators.

Finally, future researchers could examine the learning outcomes of the peer-generated tests. Given the nature of the tests, it is likely if not certain that some learning occurs as a result of taking the test, but to what extent is unknown. In addition, how much learning, if any, occurs as a result of administering the test could be investigated.

Final Comments

The findings of this study support the use of word cards as a deliberate vocabulary learning technique and suggest that regular testing can be beneficial, especially if the tests are designed in such a way as to increase intrinsic motivation.

In addition, while the self-selected and self-made word card approach seems to offer the best possible learning outcomes, it might not be popular with students, and ultimately the choice of which method is most appropriate can only be made by teachers who know their students and understand the constraints in which they are operating. For example, if a teacher is confident that they have highly motivated students, then allowing them to choose any of the methods described here seems appropriate. However, when dealing with demotivated learners and when giving students autonomy in choosing a deliberate study method is not a viable option, choosing the self-selected and self-made approach might be best because it maximizes involvement and offers very good potential learning outcomes.

The tests investigated in this study have already been used and positively evaluated in a number of contexts by several EFL teachers, but now there is substantial evidence that the tests increase motivation and can be used as a reliable measure of learners' vocabulary learning and the effort they are putting into acquiring new vocabulary. If teachers are responsible for groups of learners who are all studying the same target vocabulary sets, then the tests might not be needed for assessment purposes, but they still offer an enjoyable and motivating way for learners to review and receive feedback on their vocabulary knowledge and progress.

Overall, the results of the three experiments indicate that (a) word cards are an effective and popular deliberate vocabulary learning method, and (b) regular peer-to-peer testing can increase motivation and provide learners with much needed and appreciated feedback on their vocabulary development.

REFERENCES

Abrams, S. S., & Walsh, S. (2014). Gamified vocabulary. *Journal of Adolescent & Adult Literacy, 58*(1), 49–58. doi:10.1002/jaal.315

Adapa, S. (2015). Integrating teaching resources and assessment tasks to enhance student experience. *International Journal of Learning, Teaching and Educational Research, 10*(4), 28–39. Retrieved from www.ijlter.org/index.php/ijlter/article/viewFile/287/pdf

Agawa, G., Black, G., & Herriman, M. (2011). Effects of web-based vocabulary training for TOEIC. In A. Stewart (Ed.), *JALT2010 Conference Proceedings* (pp. 160–167). Tokyo: JALT. Retrieved from http://jalt-publications.org/proceedings/articles/1099-effects-web-based-vocabulary-training-toeic

Al-Murtadha, M. (2013). Investigating the impact of sentence writing on second language vocabulary acquisition. *Vocabulary Education and Research Bulletin, 2*(2), 2–3. Retrieved from http://jaltvocab.weebly.com/publications.html

Al-Murtadha, M. (2014). The relationship between productive vocabulary tasks and accuracy, fluency, and motivation of Japanese EFL learners. *Vocabulary Education and Research Bulletin, 3*(1), 2–4. Retrieved from http://jaltvocab.weebly.com/publications.html

Altiner, C. (2011). *Integrating a computer-based flashcard program into academic vocabulary learning.* (Master's Thesis, Iowa State University). Retrieved from https://lib.dr.iastate.edu/etd/10160/

Anderson, J. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching*: *Research reviews* (pp. 77–177). Newark, DE: International Reading Association.

Astika, G. G. (1993). Analytical assessment of foreign students' writing. *RELC Journal*, *24*(1), 61–72. doi:10.1177/003368829302400104

Ballance, O., & Cobb, T. (2020). Resources for learning single-word items. In S. Webb (Ed)*, The Routledge handbook of vocabulary studies* (pp. 320–335). New York, NY: Routledge.

Bamford, J., & Day, R. R. (1998). Teaching reading. *Annual Review of Applied Linguistics*, *18*, 124–141. doi:10.1017/s0267190500003512

Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *Modern Language Journal, 87*(4), 546–561. doi:10.1111/1540-4781.00207

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research*, *20*(4), 303–334. doi:10.1191/0267658304sr233oa

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research, 22*(4), 487–497. doi:10.1191/0267658304sr233oa

Barcroft, J. (2007a). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35–56. doi:10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2007b). Effects of word and fragment writing during l2 vocabulary learning. *Foreign Language Annals, 40*(4), 713–726. doi:10.1111/j.1944-9720.2007.tb02889.x

Barcroft, J. (2009). Strategies and performance in intentional L2 vocabulary learning. *Language Awareness, 18*(1), 74–89. doi:10.1080/09658410802557535

Barcroft, J. (2015). *Lexical input processing and vocabulary learning.* Amsterdam, The Netherlands: Benjamins.

Behzadi, A., & Haji Pour Nezhad, G. (2014). The effect of two task types on learning English words and idioms. *International Journal of English Language and Literature Studies*, *3*(1), 12–20. Retrieved from http://www.aessweb.com/pdf-files/ijells 3(1), 12–20.pdf

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*(1), 15–32. doi:10.1111/j.1467-9817.1984.tb00252.x

Blanca, M. J., Alarcón, R., & Arnau, J. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema 29*(4), 552–557. doi:10.7334/psicothema2016.383

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction*, *1*(1), 20–28. doi:10.7820/vli.v01.1.brown

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(2), 1–10. doi:10.7820/vli.v03.2.2187-2759

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. http://dx.doi.org/10.1037/a0019902

Carter, R., & McCarthy, M. (2014). Lexis and discourse: Vocabulary in use. In R. Carter, & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 201–220). London, England: Routledge.

Chow, S. J. (2007). *An investigation of vocabulary coverage in English language teaching textbooks used in Hong Kong secondary schools* (Master's Thesis, University of Hong Kong, Pokfulam, Hong Kong). Retrieved from http://hub.hku.hk/handle/10722/51736

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

Conner, R. F. (1980). Ethical issues in the use of control groups. *New Directions for Evaluation, 7*, 63–75. doi:10.1002/ev.1253

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–218. doi:10.2307/3587951

Coxhead, A. (Ed.) (2014). *New ways in teaching vocabulary, revised*. Alexandria, VA: TESOL.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. Retrieved from https://www.sciencedirect.com/science/article/pii/S002253717280001X

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. Retrieved from http://psycnet.apa.org/record/1976-00185-001

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language, 7*(2), 541–551. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl72day.pdf

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18,* 105–115. doi:10.1037/h0030644

Deci, E. L. (1992). The relation of interest to the motivation of behavior: a self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 43–70). Hillsdale, NJ: Erlbaum.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Dodigovic, M. (2013). Vocabulary learning with electronic flashcards: Teacher design vs. student design. *Voices in Asia Journal, 1*, 15–33. Retrieved from www.viajournal.org/admin/file_upload_name/pp._15-33.pdf

Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge, England: Cambridge University Press.

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research, 16*(2), 227–252. doi:10.1177/1362168811431377

Edy, S. (2014). The effectiveness of extensive reading on students' reading comprehension achievement as observed from students' motivation. *Education Journal of Social Sciences, 2*(1), 54–58. doi:10.17977/jph.v2i1.4443

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

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*(2)*,* 174–187. doi:10.2307/747863

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning*, *41*(3), 375–411. doi:10.1111/j.1467-1770.1991.tb00611.x

Elley, W. B., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly, 19*(1), 53–67. doi:10.2307/747337

Ellis, N. C. (1994). Introduction: implicit and explicit language learning—An overview. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1–31). London, England: Academic Press.

Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning, 8*(2), 1031–28. doi:10.1080/0958822940080202

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143–188. doi:10.1017.S0272263102002024

Ellis, N. C. (2015). Implicit and explicit learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 3–23). Amsterdam, The Netherlands: Benjamins.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning*, *54*(2), 227–275. doi:10.1111/j.1467-9922.2004.00255.x

Emrah, B., & Omur, A. (2010). A comparison of undergraduate students' English vocabulary learning: Using mobile phones and flash cards. *The Turkish Online Journal of Educational Technology, 9*(3), 1–7. Retrieved from http://files.eric.ed.gov/fulltext/EJ898010.pdf

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. doi:10.1007/bf00197068

Farrokhi, F., & Mahmoudi-Hamidabad, A. (2012). Rethinking convenience sampling: Defining quality criteria. *Theory and Practice in Language Studies, 2*(4), 784–792. doi:10.4304/tpls.2.4.784–792

Felix, U. (2008). The unreasonable effectiveness of CALL: What have we learned in two decades of research? *ReCALL, 20*(2), 141–161. doi:10.1017/s0958344008000323

Fendler, L. (2016). Ethical implications of validity-vs.-reliability trade-offs in educational research. *Ethics and Education*, *11*(2), 214–229. doi:10.1080/17449642.2016.1179837

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). London, England: Sage.

Folse, K. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*(2), 273–293. doi:10.2307/40264523

Folse, K. (2011). Applying L2 lexical research findings in ESL teaching. *TESOL Quarterly*, *45*(2), 362–369. doi:10.5054/tq.2010.254529

French, E. G. (1955). Some characteristics of achievement motivation. *Journal of Experimental Psychology, 50*(4), 232–236.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* London, England: Routledge.

Geis, G. L. (1984). Comparing instructional methods: Some basic research problems. *Canadian Journal of Higher Education, 14*(2), 91–98. Retrieved from http://journals.sfu.ca/cjhe/index.php/cjhe/article/viewFile/182937/182918

Hadfield, J., & Dörnyei, Z. (2013). *Motivating learning*. Harlow, England: Pearson.

Haggard, M. R. (1985). An interactive strategies approach to content reading. *Journal of Reading, 29*, 204–210. Retrieved from http://www.jstor.org/stable/40029660?seq=1#page_scan_tab_contents

Haggard, M. R. (1986). The vocabulary self-collection strategy: Using student interest and world knowledge to enhance vocabulary growth. *Journal of Reading, 26*, 634–642. Retrieved from http://www.jstor.org/stable/40029691?seq=1#page_scan_tab_contents

Han, Z. H. & Cheng-ling, A. (2010). Repeated-reading-based instructional strategy and vocabulary acquisition: A case study of a heritage speaker of Chinese. *Reading in a Foreign Language, 22*(2), 242–262. Retrieved from http://nflrc.hawaii.edu/rfl/October2010/articles/han.pdf

Harackiewicz, J. M., & Sansone, C. (2000). Rewarding competence: The importance of goals in the study of intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 79–104). San Diego, CA: Academic Press

Harmon, J., Wood, K., Hedrick, W., & Gress, M. (2008). "Pick a word—Not just any word". Using vocabulary self-selection with expository texts. *Middle School Journal, 40*, 43–52. Retrieved from http://eric.ed.gov/?id=EJ809642

Hatch, J. A. (2002). *Doing qualitative research in education settings.* Albany, NY: State University of New York Press

Hazenberg, S., & Hulstijn J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, *17*(2), 145–163. doi:10.1093/applin/17.2.145

Heidari-Shahreza, M. A., & Tavakoli, M. (2012). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition by Iranian EFL learners. *The Language Learning Journal, 44*(1), 17–32. doi:10.1080/09571736.2012.708051

Hirai, A. (2014). A review of four studies on measuring vocabulary knowledge. *Vocabulary Learning and Instruction, 3*(2), 85–92. doi:10.7820/vli.v03.2.hirai

Hirschel, R., & Fritz, E. (2013). Learning vocabulary: CALL program versus vocabulary notebook. *System, 41*(3), 639–653. doi:10.1016/j.system.2013.07.016

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–696. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl82hirsh.pdf

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207–223. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl112horst.pdf

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403–430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 21*(2), 181–193. doi:10.1017/s0272263199002028

Hulstijn, J. H. (1992). Retention of inferred and given words meanings: Experiments in incidental vocabulary learning. *Vocabulary and Applied Linguistics*, 113–125. doi:10.1007/978-1-349-12396-4_11

Hulstijn J.H. (1992) Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P.J.L., Arnaud, & H., Béjoint (Eds.),V*ocabulary and applied linguistics* (pp. 113–125). London, England: Palgrave Macmillan

Hulstijn, J. H. (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., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80*(3), 327–339. doi:10.1111/j.1540-4781.1996.tb01614.x

Hulstijn, J., & Trompetter, P. (1998). Incidental learning of second language vocabulary in computer-assisted reading and writing tasks. In D. Albrechtsen, B. Henrikse, I. M. Mees, & E. Poulsen (Eds.), *Perspectives on foreign and second language pedagogy* (pp. 191–200). Odense, Denmark: Odense University Press.

Hung, H. T. (2015). Intentional vocabulary learning using digital flashcards. *English Language Teaching, 8*(10), 107–112. doi:10.5539/elt.v8n10p107

Iso, T.(2012). Examining the validity of the lexical access time test (LEXATT2). *Vocabulary Learning and Instruction, 1*(1), 78–82. doi:10.7820/vli.v01.1.iso

Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar based on the LOB Corpus,* (Vols. 1–2). Oxford, England: Clarendon Press.

Kanayama, K., & Kasahara, K. (2018). The indirect effects of testing: Can poor performance in a vocabulary quiz lead to long-term l2 vocabulary retention? *Vocabulary Learning and Instruction, 7*(1), 1–13. doi: 10.7820/vli.v07.1.2187-2759

Kaneko, M. (2013). Lexical frequency profiling of Japanese high school English reading textbooks. *Vocabulary Education and Research Bulletin, 2*(1), 3–4. Retrieved from http://jaltvocab.weebly.com/publications.html

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319, 966–968. doi: 10.1126/science.1152408

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. doi:10.1177/1362168808089922

Khan, A., & Rayner, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem. J*ournal of Applied Mathematics and Decision Sciences, 7*(4), 187–206. doi:10.1155/s1173912603000178

Kitano, M., & Chiba, K. (2019). Comparing the effectiveness of word cards and list learning with Japanese learners of English. *Vocabulary Learning and Instruction, 8*(1), 70–75. doi:10.7820/vli.v08.1.kitano.chiba

Kojic-Sabo, I., & Lightbown, P. M. (1999). Students’ approaches to vocabulary learning and their relationship to success. *Modern Language Journal, 83*(2), 176–192. doi:10.1111/0026-7902.00014

Komachali, M. E., & Khodareza, M. (2012). The effect of using vocabulary flash card on Iranian pre-university students' vocabulary knowledge. *International Education Studies*, *5*(3), 134–147. doi:10.5539/ies.v5n3p134

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464. doi:10.1111/j.1540-4781.1989.tb05325.x

Kremmel, B., & Schmitt, N. (2017). Vocabulary Levels Test. *The TESOL Encyclopedia of English Language Teaching*, 1–6. doi:10.1002/9781118784235.eelt0499

Kuo, Y., & Ho, H. Y. (2012). Effects of word card strategy versus word list strategy on Taiwanese EFL junior high school students' vocabulary retention. *Electronic Journal of Foreign Language Teaching*, *9*(1), 26–45. Retrieved from http://e-flt.nus.edu.sg/v9n12012/kuo.pdf

Laerd Statistics. (2018). *MANOVA and ANOVA in SPSS statistics.* Retrieved from https://statistics.laerd.com/spss-tutorials-using-spss-statistics.php

Laerd Statistics. (2019). *Advanced guide to statistics in SPSS statistics.* Retrieved from https://statistics.laerd.com/

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, England: Academic Press. Retrieved from http://bjorklab.psych.ucla.edu/pubs/Landauer.Bjork.1978.pdf

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Laurén & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, England: Multilingual Matters.

Laufer, B. (1992a). How much lexis is necessary for reading comprehension? In: P.J.L., Arnaud, & H., Béjoint (Eds.),*Vocabulary and applied linguistics* (pp. 126–132). London, England: Palgrave Macmillan

Laufer, B. (1992b). Reading in a foreign language: How does L2 lexical knowledge Interact with the learner's general academic ability? *Journal of Research in Reading*, *15*, 126–132. doi:10.1111/j.1467-9817.1992.tb00025.x

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review/LaRevue canadienne des langues vivantes, 59*, 567–587. doi:10.3138/cmlr.59.4.567.

Laufer, B. (2014). Vocabulary in a second language: Selection, acquisition, and testing: A commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction, 3*(2), 38–46. doi:10.7820/vli.v03.2.laufer

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436. doi:10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. doi:10.1093/applin/22.1.1

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. doi:10.1093/applin/16.3.307

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited. *Reading in a Foreign Language, 22*(1), 15–30. Retrieved from http://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Laufer, B., & Rozovski-Roitblat, B. (2014). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research, Special Issue: Vocabulary Research and Pedagogy,* 1–25. doi:10.1177/1362168814559797

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108. doi:10.1177/003368829702800106

Lavoie, C. (2016). The effect of training on vocabulary strategy use: Explicit teaching of word family, word network and word card strategies. *The Journal of Language Teaching and Learning, 6*(1), 20–34. Retrieved from www.jltl.org/index.php/jltl/article/download/454/113

Liu, Y. (2011). *Vocabulary recognition and memorization: A comparison of two methods.* (Masters' Thesis, Kristianstad University, Kristianstad, Sweden). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A429600&dswid=-9232

Luppescu, S., & Day, R. R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning, 43*(2), 263–279. doi:10.1111/j.1467-1770.1992.tb00717.x

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.

Mantiri, O. (2015). Principles in language learning motivation. *Linguistics and Literature Studies, 3*(2), 70–74. doi:10.13189/lls.2015.030206

Masoudi, G. (2017). The effect of vocabulary self-selection strategy and input enhancement strategy on the vocabulary knowledge of Iranian EFL learners. *English Language Teaching*, *10*(8), 32–42. doi:10.5539/elt.v10n8p32

Masson, M. (2012). Student feedback regarding the use of 'mastery sentences'. *Vocabulary Education and Research Bulletin, 1*(1), 5–6. Retrieved from http://jaltvocab.weebly.com/publications.html

Mcdaniel, M. A., Roediger, H. L., & Mcdermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200–206. doi:10.3758/bf03194052

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. doi:10.7820/vli.v03.2.mclean.et.al

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin and Review, 14*, 225–229. doi:10.3758/BF03194056

Mizumuoto, A. (2010). Exploring the art of vocabulary learning strategies: A closer look at Japanese EFL university students. Tokyo, Japan: Kinseido. Retrieved from http://www.mizumot.com/files/book2010

Mizumoto, A. (2013). Enhancing self-efficacy in vocabulary learning: A self-regulated learning approach. *Vocabulary Learning and Instruction, 2*(1), 15–24. doi:10.7820/vli.v02.1.mizumoto

Moir J., & Nation I. S. P. (2002). Learners' use of strategies for effective vocabulary learning. *Prospect, 17*(1), 15–35. Retrieved from http://www.ameprc.mq.edu.au/docs/prospect_journal/volume_17_no_1/17_1_2_Moir.pdf

Mondria, J. A., & Boer, M. W. D. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics, 12*(3), 249–267. doi:10.1093/applin/12.3.249

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. doi:10.1016/0346-251x(94)90039-6

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 79–100). Amsterdam, The Netherlands: Benjamins.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233–253. doi: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*(1), 3–20. doi:10.1017/s09583440080002

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,* 17–38. doi:10.1080/09588221.2010.520675

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? S*tudies in Second Language Acquisition, 37*, 677–711. doi:10.1017/S0272263114000825

Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed), *The Routledge handbook of vocabulary studies* (pp. 304–319). New York, NY: Routledge.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12–25. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82. doi:10.3138/cmlr.63.1.59

Nation, I. S. P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching, 1*(1), 2–13. doi:10.2167/illt039.0

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, England: Cambridge University Press.

Nation, 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. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage.

Newton, J. (2013). Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research, 17,* 164–187. doi:10.1177/1362168812460814

Nikoopour, J., & Kazemi, A. (2014). Vocabulary learning through digitized & non-digitized flashcards delivery. *Procedia—Social and Behavioral Sciences, 98*, 1366–1373. doi:10.1016/j.sbspro.2014.03.554

Ng, C. F., & Ng, P. K. (2015). A review of intrinsic and extrinsic motivations of ESL learners. *International Journal of Languages, Literature and Linguistics, 1*(2), 98–105. doi:10.7763/ijlll.2015.v1.20

Nonaka, T. (2004). English vocabulary size of Japanese university students. *Niigata Seiryo University Junior College Research Report, 34*, 25–33. Retrieved from www.n-seiryo.ac.jp/library/kiyo/tkiyo/04pdf/T0403.pdf

Oberg, A. (2011). Comparison of the effectiveness of a CALL-based approach and a card-based approach for vocabulary acquisition and retention. *CALICO Journal, 29*(1), 118–144.

Ogawa, K. (2012). *EFL learner autonomy and unfamiliar vocabulary learning* (Doctoral Dissertation, Temple University, Tokyo, Japan). Retrieved fromhttp://digital.library.temple.edu/cdm/ref/collection/p245801coll10/id/174127

O'Leary, K. D., & Borkovec, T. D. (1978). Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist, 33*(9), 821–830. http://dx.doi.org/10.1037/0003-066X.33.9.821

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing: An Interdisciplinary Journal*, *22*(5), 545–565. doi:10.1007/s11145-008-9124-z

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language, 22*(1)*,* 31–55. Retrieved from http://nflrc.hawaii.edu/rfl/April2010/articles/pellicersanchez.pdf

Pichette, F., de Serres, L., & Lafontaine, M. (2011). Sentence reading and writing for second language vocabulary acquisition. *Journal of Applied Linguistics, 32*(5), 1–18. doi:10.1093/applin/amr037

Pimsleur, P. (1967). A memory schedule. *Modern Language Journal*, *51*(2), 73–75.

doi:10.1111/j.1540-4781.1967.tb06700.x

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*(2), 282–308. doi:10.3138/cmlr.56.2.282

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536. doi:10.1111/1467-9922.00193

Reynolds, B. L., & Shih, Y. C. (2019). The learning effects of student-constructed word cards as homework for the adolescent English Language classroom. *System*, *81*, 146–162. doi:10.1016/J.SYSTEM.2019.01.005

Resendes, M., Chen, B., Acosta, A., & Scardamalia, M. (2013). The effect of formative feedback on vocabulary use and distribution of vocabulary knowledge in a grade two knowledge building class. *CSCL (1)*, 391–405. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.725.781&rep=rep1&type=pdf

Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching, 47*(2), 135–173. doi:10.1017/S0261444813000505

Rieder, A. (2003). Implicit and explicit learning in incidental vocabulary acquisition. *Vienna English Working Papers, 12*(2). Retrieved from http://www.univie.ac.at/Anglistik/views/03_2/RIE_SGLE.PDF

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Rott, S., & Williams, J. (2003). Making form-meaning connections while reading: A qualitative analysis of word processing. *Reading in a Foreign Language*, *15(*1), 45–75.

Ruddell, M. R., & Shearer, B. A. (2002). “Extraordinary,” “tremendous,” “exhilarating,” “magnificent”: Middle school at-risk students become avid word learners with the vocabulary self-collection strategy (VSS). *Journal of Adolescent & Adult Literacy*, *45*, 352–363.

Ruegg, R. (2013). A comparison of lexical feedback on writing received from peers and a teacher. *Vocabulary Learning and Instruction, 2*(1), 25–31. doi:10.7820/vli.v02.1.ruegg

Ruegg, R., & Brown, C. (2014). Analyzing the effectiveness of textbooks for vocabulary retention. *Vocabulary Learning and Instruction, 3*(2), 11–18. doi:10.7820/vli.v03.2.ruegg.brown

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78. doi:10.1037/0003-066x.55.1.68

Sakai, H., & Kikuchi, K. (2009). An analysis of demotivators in the EFL classroom. *System, 37*(1), 57–69. doi:10.1016/j.system.2008.09.005

Sarbazi, M. R. (2014). Involvement load hypothesis: Recalling unfamiliar words meaning by adults across genders. *Procedia-Social and Behavioral Sciences, 98*, 1686–1692. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877042814026858

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. (2000). *Vocabulary in language teaching*. Cambridge, England: Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329–363. doi:10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, England: Palgrave Macmillan.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212–226. doi:10.1017/S0261444815000075

Schmitt, N., & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal, 49*(2), 133–143. doi: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. doi:10.1191/026553201668475857

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction, 1*(1), 53–59. doi:10.7820/vli.v01.1.stewart

Tanaka, M. (2017). Examining EFL vocabulary learning motivation in a demotivating learning environment. *System*, *65*, 130–138. doi:10.1016/j.system.2017.01.010

Thompson, A. (2014). Measuring university student interest in vocabulary. *Vocabulary Education and Research Bulletin, 3*(2), 15–16. Retrieved from http://jaltvocab.weebly.com/publications.html

Tseng, W. T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357–400. doi:10.1111/j.1467-9922.2008.00444.x

Tsubaki, M. (2012). The involvement load hypothesis: An inquiry into vocabulary learning. *Journal of Bunkyo Gakuin University, Department of Foreign Languages and Bunkyo Gakuin College, 6*, 175–184. Retrieved from http://www.cicero.u-bunkyo.ac.jp/lib.

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage, 78*, 94–102. doi:10.1016/j.neuroimage.2013.03.071

van Zeeland, H. & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics 34*(4), 457–479. doi:10.1093/applin/ams074

Verma, J. P. (2016). Two-factor mixed MANOVA. Retrieved from https://www.slideshare.net/JPVerma3/twofactor-mixed-manova-with-spss

Vidal, K. (2003). Academic listening: A source of vocabulary acquisition? *Applied Linguistics*, *24*(1), 56–89. doi:10.1093/applin/24.1.56

Walqui, A. (2000). Contextual factors in second language acquisition. *The Educational Resources Information Center (ERIC) Digest*. Retrieved from https://eric.ed.gov/?id=ED444381

Walters, J. D., & Bozkurt, N. (2009). The effect of keeping vocabulary notebooks on vocabulary acquisition. *Language Teaching Research, 13*(4), 403–423. doi:10.1177/1362168809341509

Waring, R. (1997). A study of receptive and productive learning from word cards. *Studies in Foreign Languages and Literature*, *21*, 94–114. Retrieved from http://www.robwaring.org/papers/various/wordcard.html

Waring, R. (2004) In defence of learning words in word pairs: But only when doing it the 'right' way! Retrieved from http://www1.harenet.ne.jp/~waring/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/October2003/waring/waring.html

Watanabe, Y. (1997). Input, intake, and retention: Effects of increasing processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*(3), 287–307. doi:10.1017/s027226319700301x

Webb, S. (2002). *Investigating the effects of learning tasks on vocabulary knowledge* (Doctoral Dissertation, Victoria University of Wellington, Wellington, New Zealand). Retrieved from https://researcharchive.vuw.ac.nz/xmlui/handle/10063/612

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33–52. doi:10.1017/S0272263105050023

Webb, S. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research, 11*(1), 63–81. doi:10.1177/1362168806072463

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal, 40*(3), 360–376. doi:10.1177/0033688209343854

Webb, S. (2020). Introduction. In S. Webb (Ed), *The Routledge handbook of vocabulary studies* (pp. 1–12). New York, U.S.A.: Routledge.

Webb, S., & Piasecki, A. (2018). Re-examining the effects of word writing on vocabulary learning. *ITL-International Journal of Applied Linguistics, 169*(1), 72–93. doi:10.1075/itl.00007.web

Wegener, D., T., & Blankenship, K., L. (2007). Ecological validity. In R. F. Baumeister & K. D. Vohs, (Eds), *Encyclopedia of social psychology* (p. 167). Thousand Oaks, CA: Sage Publications. doi:10.4135/9781412956253

Wilkins, D. (1972). *Linguistics in language teaching*. London, England: Arnold.

Wilkinson, D. (2015). Peer testing to increase motivation for deliberate vocabulary study. *Vocabulary Education and Research Bulletin, 4*(2), 2–4. Retrieved from http://jaltvocab.weebly.com/publications.html

Wilkinson, D. (2017). EFL Vocabulary acquisition through word cards: Student perceptions and strategies. *TESL-EJ*, *21*(3). Retrieved from http://www.tesl-ej.org/wordpress/issues/volume21/ej83/ej83a4/

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal*, *40*(1), 125–137. doi:10.1080/09571736.2012.658232

Workman, E. A., & Williams, R. L. (1980). Effects of extrinsic rewards on intrinsic motivation in the classroom. *Journal of School Psychology, 18*(2), 141–147. doi:10.1016/0022-4405(80)90029-1

Yoshii, M. (2014). Effects of glosses and reviewing of glossed words on L2 vocabulary learning through reading. *Vocabulary Learning and Instruction, 3*(2), 19–30. doi:10.7820/vli.v03.2.yoshii

Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics, 35*(3), 283–304. doi:10.1093/applin/amt014

Zheng, Y. (2012). Exploring long-term productive vocabulary development in an EFL context: The role of motivation. *System*, *40*(1), 104–119. doi:10.1016/j.system.2012.01.007

APPENDICES

APPENDIX A

FINAL WORD SELECTION AFTER TRANSLATION AND PILOTING

	English	Part of Speech	Final Translation Choice
1.	inventory	n	在庫リスト
2.	civic	adj	市民の
3.	dynamic	adj	力強い
4.	prohibit	v	禁じる
5.	hay	n	干し草
6.	stumble	v	つまずく
7.	dawn	n	夜明け
8.	descend	v	降りる
9.	silly	adj	愚かな
10.	readily	adv	容易に
11.	revolutionary	adj	革命的な
12.	ruling	n	支配
13.	annually	adv	毎年
14.	undertake	v	引き受ける
15.	ridiculous	adj	ばかげた
16.	credibility	n	信頼性
17.	sentiment	n	感情
18.	alleged	adj	申し立てられた
19.	dynamics	n	動力学
20.	dignity	n	尊厳
21.	tuck	v	押し込む
22.	dilemma	n	窮地
23.	liability	n	責任
24.	merchant	n	商人
25.	ancestor	n	祖先
26.	condemn	v	非難する
27.	optimistic	adj	楽観的な
28.	statute	n	法
29.	conceive	v	思いつく
30.	haul	v	運搬する
31.	concede	v	認める
32.	spectacular	adj	壮観な
33.	immune	adj	免疫の
34.	ideological	adj	観念的な
35.	secular	adj	世俗的な
36.	manipulate	v	操作する
37.	utilize	v	活用する

38. presumably	adv	おそらく
39. consequently	adv	その結果として
40. inherit	v	相続する
41. envision	v	想像する
42. rational	adj	合理的
43. exclusively	adv	もっぱら
44. terrain	n	地形
45. deployment	n	配置
46. migration	n	移住
47. wound	v	負傷させる
48. coup	n	政変
49. stimulus	n	刺激
50. sphere	n	球体
51. boast	v	自慢する
52. excessive	adj	過度の
53. deliberately	adv	意図的に
54. striking	adj	目立つ
55. isolated	adj	孤立した
56. demographic	adj	人口統計の
57. jurisdiction	n	管轄
58. inevitably	adv	必然的に
59. constraint	n	制約
60. expedition	n	探検隊
61. allegedly	adv	申し立てによると
62. solely	adv	もっぱら
63. formerly	adv	以前は
64. widow	n	未亡人
65. autonomy	n	自治
66. routinely	adv	日常的に
67. fragment	n	断片
68. yield	n	収穫高
69. ego	n	自我
70. rehabilitation	n	社会復帰
71. confession	n	白状
72. trauma	n	心的外傷

APPENDIX B

LEARNER DICTIONARY BOOKLET

English	Part of speech	Japanese	Example sentence
inventory	N	在庫リスト	The bookstore manager did an **inventory** of the books, so she wouldn't run out of the most popular ones.
prohibit	V	禁じる	Soka university **prohibits** smoking anywhere on campus.
hay	N	干し草	It is important to have lots of **hay** to feed cows in winter because there is no fresh grass.
stumble	V	つまずく	While walking home in the dark I stepped on something and **stumbled**, but I didn't fall over.
dawn	N	夜明け	I have so much homework I wake up at **dawn** every morning to start doing it.
descend	V	降りる	When arriving at its destination, it takes an airplane about 30 minutes to **descend** to the ground.
readily	Adv	容易に	Unlike many countries in Africa, fresh water is **readily** available in Japan.
revolutionary	Adj	革命的な	Many people think the iPhone was a **revolutionary** piece of technology.
ruling	N	支配	The judge's **ruling** to fine (charge) the company 1 million US$ surprised many people.
undertake	V	引き受ける	My father said "**Undertake** something that is difficult; it will do you good"
ridiculous	Adj	ばかげた	The student told the teacher a **ridiculous** story about her dog eating her homework.
credibility	N	信頼性	The student lost all **credibility** after he was caught lying and cheating on his test.
sentiment	N	感情	I think I can really understand the painter's **sentiment** when I look at his painting.
alleged	Adj	申し立てられた	The **alleged** criminal has said many times that he is innocent (did not commit the crime).

dynamics	N	動力学	Group **dynamics** are important when choosing teams; you must think if the people can work together well.
dignity	N	尊厳	It is important for children in sport to learn to both win and lose with **dignity**
tuck	V	押し込む	My father **tucks** his trousers into his socks to keep them dry when it's raining
dilemma	N	窮地	I have a **dilemma**: I want the job I was offered because the money is great, but I think I will hate it.
liability	N	責任	The company has admitted **liability** for the accident, and is paying all medical costs to the victims.
merchant	N	商人	There are now a small number of specialist wine **merchant**s which only sell online.
ancestor	N	祖先	Although my parents and I are British, my **ancestors** come from Denmark.
condemn	V	非難する	Most people would **condemn** the sale of cigarettes to young children.
optimistic	Adj	楽観的な	I am **optimistic** about my test results; I think I did well.
statute	N	法	Our town has a **statute** that stops street parking between midnight and six AM.
haul	V	運搬する	Ships are used to **haul** heavy goods between countries.
concede	V	認める	I must **concede** that math is not my strong subject.
rational	Adj	合理的	It is not **rational** to expect a high grade if you are often late and don't do homework.
exclusively	Adv	もっぱら	I work **exclusively** for Soka University; I do not work at any other universities.
terrain	N	地形	America many kinds of **terrain**: deserts, mountains, oceans, forests, etc.
deployment	N	配置	The **deployment** of American soldiers to Iraq was not popular with many people.

APPENDIX C

L1-L2 TEST: ACTIVE RECALL TEST

	Japanese	English
1	顕著	
2	革命的な	
3	非難する	
4	降りる	
5	配置	
6	過度の	
7	運搬する	
8	責任	
9	負傷させる	
10	認める	
11	自慢する	
12	窮地	
13	移動	
14	禁じる	
15	祖先	
16	申し立てられた	
17	球体	
18	法	
19	楽観的な	
20	政変	
21	支配	
22	押し込む	
23	感情	
24	意図的に	
25	引き受ける	
26	干し草	
27	尊厳	
28	容易に	
29	夜明け	
30	地形	
31	在庫リスト	
32	商人	
33	合理的	
34	司法権	

35	動力学	
36	刺激	
37	信頼性	
38	もっぱら	
39	ばかげた	
40	つまずく	

APPENDIX D

L2-L1 TEST: PASSIVE RECALL TEST

	English	Japanese
1	alleged	
2	ancestor	
3	boast	
4	concede	
5	condemn	
6	coup	
7	credibility	
8	dawn	
9	deliberately	
10	deployment	
11	descend	
12	dignity	
13	dilemma	
14	dynamics	
15	excessive	
16	exclusively	
17	haul	
18	hay	
19	inventory	
20	jurisdiction	
21	liability	
22	merchant	
23	migration	
24	optimistic	
25	prohibit	
26	rational	
27	readily	
28	revolutionary	
29	ridiculous	
30	ruling	
31	sentiment	
32	sphere	
33	statute	
34	stimulus	
35	striking	
36	stumble	
37	terrain	

38	tuck	
39	undertake	
40	wound	

APPENDIX E

L2 CLOZE TEST

1. The bookstore manager did an **in**____________of the books before semester started.
2. Many people agree that it is a good idea to **pr**____________ smoking on campus.
3. Farmers feed cows **h**____________ in winter because there is no fresh grass.
4. While walking home I stepped on something and **st**__________d, but didn't fall over.
5. I have so much homework I wake up at **d**____________ every morning to study.
6. It takes an airplane about 30 minutes to **de**____________ to the ground during landing.
7. Unlike many countries in Africa, fresh water is **r**____________ available in Japan.
8. Many people think that the iPhone was a **re**____________ piece of technology.
9. The judge's **r**____________to fine the company 1 million dollars surprised many people.
10. My father said "**un**____________something that is difficult; it will do you good".
11. The student told the teacher a **ri**____________ story about her dog eating her homework.
12. The student lost all **cr**____________ after he was caught cheating on his test.
13. Watching her daughter getting married brought back the **s**____________ of her own wedding day.
14. The **al**____________ criminal says that he is innocent (did not commit the crime).
15. Group **d**____________ are important when choosing teams because people have different personalities.
16. When playing sports, children must learn to win and lose with **d**____________.
17. I **t**____________ my trousers into my socks to keep them dry when it rains.
18. My **d**____________ is that the salary is very high, but the work is boring.
19. The company admitted **l**____________ for the accident, and will pay all the medical costs.
20. I found a really good wine **m**____________ who delivers to my house for free.

21. Although my parents and I are British, my **a**___________**s** come from Germany.

22. Most people **c**___________ the sale of cigarettes and alcohol to young children.

23. I am very **o**___________ about my test results; I think I did very well.

24. The town **s**___________ stops parking on the street between 9 am and 6 pm.

25. Very large container ships are used to **h**___________ heavy goods between countries.

26. I must **c**___________ that math and science are not my strongest subjects.

27. It is not **r**___________ to expect a high grade if you never do homework.

28. I work **e**___________ for this university; I do not work at any other universities.

29. America has many kinds of **t**___________, for example, deserts, mountains, and forests.

30. The **d**___________ of American soldiers to Okinawa is not popular with many Japanese people.

31. There has been a lot of **m**___________ recently from Africa and Syria to Europe.

32. It is important to clean a **w**___________ carefully so it does not get infected.

33. There was a **c**__________ in Thailand recently when the army took over government control.

34. Many people thing that Japan's economic **s**___________ plan is not working very well.

35. People thought the world was flat, but now we know it is a **s**___________.

36. It is not polite to **b**___________ about the things you are good at.

37. Everyone knows that **e**___________ drinking of alcohol is very bad for our health.

38. It is important to **d**___________ study vocabulary if you want to improve your English.

39. Thanks to satellites, there are many **s**___________ images of the earth from space.

40. In America, no state has **j**___________ over the others; they all have self-control.

APPENDIX F

INTERVIEW PROTOCOL

Research focus	Interview questions
Word card method	What do you think about word cards in general? What do you think about making word cards? How effective do you think it is/they are as a way to learn/study vocabulary? What do you think of word cards in terms of efficiency? What do you think the benefits are of using word cards? Why/How do they help you learn vocabulary? How do they compare to other ways that you have studied vocabulary (in terms of efficiency, ability to memorize etc.)?
Time-on-task and Learner Strategies	How long do you spend studying from word cards a week? How many times/sessions do you usually have? How long do you spend each time? How long does it take to make the cards? How do you study from word cards? What do you do? How do you organize your cards/piles? Do you shuffle your pack? Which side do you start with (JP-Eng or Eng-JP)? How often do you look at "known words"?
Perceptions of the three word card methods	How do you feel about each of the word card methods you used? What are some advantages and disadvantages of each method? If it was your choice, which method would you like to use?
Perceptions of the test system	How do you feel about the test system? How would you rate the difficulty of the tests? How important are your grades on the test? What are the good and bad points of the test? Does having the test affect your study habits?

APPENDIX G

PREMADE WORD CARDS (BEFORE DUPLEX PRINTING AND CUTTING)

Inventory (N)	在庫リスト(N) The bookstore manager did an **inventory** of the books, so she wouldn't run out of the most popular ones.
Prohibit (V)	禁じる (V) Soka university **prohibits** smoking anywhere on campus.
Hay (N)	干し草 (N) It is important to have lots of **hay** to feed cows in winter because there is no fresh grass.
Stumble (V)	つまずく (V) While walking home in the dark I stepped on something and **stumbled**, but I didn't fall over.
Dawn (N)	夜明け (N) I have so much homework I wake up at **dawn** every morning to start doing it.
Descend (V)	降りる (V) When arriving at its destination, it takes an airplane about 30 minutes to **descend** to the ground.
Readily (Adv)	容易に (Adv) Unlike many countries in Africa, fresh water is **readily** available in Japan.
Revolutionary (Adj)	革命的な (Adj) Many people think the iPhone was a **revolutionary** piece of technology.
Ruling (N)	支配 (N) The judge's **ruling** to fine (charge) the company 1 million US$ surprised many people.
Undertake (V)	引き受ける (V)

	My father said "**Undertake** something that is difficult; it will do you good"
Ridiculous (Adj)	ばかげた (Adj) The student told the teacher a **ridiculous** story about her dog eating her homework.
Credibility (N)	信頼性 (N) The student lost all **credibility** after he was caught lying and cheating on his test.
Sentiment (N)	感情 (N) I think I can really understand the painter's **sentiment** when I look at his painting.
Alleged (Adj)	申し立てられた (Adj) The **alleged** criminal has said many times that he is innocent (did not commit the crime).
Dynamics (N)	動力学 (N) Group **dynamics** are important when choosing teams; you must think if the people can work together well.
Dignity (N)	尊厳 (N) It is important for children in sport to learn to both win and lose with **dignity**
Tuck (V)	押し込む (V) My father **tucks** his trousers into his socks to keep them dry when it's raining
Dilemma (N)	窮地 (N) I have a **dilemma**: I want the job I was offered because the money is great, but I think I will hate it.
Liability (N)	責任 (N) The company has admitted **liability** for the accident, and is paying all medical costs to the victims.
Merchant (N)	商人 (N) There are now a small number of specialist wine **merchant**s which only sell online.
Ancestor (N)	祖先 (N)

	Although my parents and I are British, my **<u>ancestors</u>** come from Denmark.
Condemn (V)	非難する (V) Most people would **<u>condemn</u>** the sale of cigarettes to young children.
Optimistic (Adj)	楽観的な (Adj) I am **<u>optimistic</u>** about my test results; I think I did well.
Statute (N)	法 (N) Our town has a **<u>statute</u>** that stops street parking between midnight and six AM.
Haul (V)	運搬する (V) Ships are used to **<u>haul</u>** heavy goods between countries.
Concede (V)	認める (V) I must **<u>concede</u>** that math is not my strong subject.
Rational (Adj)	合理的 (Adj) It is not **<u>rational</u>** to expect a high grade if you are often late and don't do homework.
Exclusively (Adv)	もっぱら (Adv) I work **<u>exclusively</u>** for Soka University; I do not work at any other universities.
Terrain (N)	地形 (N) America many kinds of **<u>terrain</u>**: deserts, mountains, oceans, forests, etc.
Deployment (N)	配置 (N) The **<u>deployment</u>** of American soldiers to Iraq was not popular with many people.

APPENDIX H

PEER-TO-PEER VERBAL VOCABULARY TEST

1. Take your partners word cards and test them on 20 of their words.
2. Ask three questions about each word.
    - How do you say ______ in Japanese/English?
    - How do you spell _______?
    - What is another form of the word?
    - Other question (your choice)

If your partner gets all 3 questions correct in the time limit = 1 point. If 1 or more incorrect = 0.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
Total										
							Total from All tests:		Average:	

APPENDIX I

PEER GENERATED WRITTEN VOCABULARY TEST

	Word in English	Word in Japanese	Part of Speech	Other Form
	1 point for each		**1/2 point for each**	**1 point for each**
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
Scores	/20		/10	/20
	Total out of 50 =		Percentage = %	

APPENDIX J

STUDENT PERCEPTIONS SURVEY QUESTIONS

1. How did you study vocabulary in high school/in other university English classes (word cards, notebooks, lists, other)?

2. A. How important do you think vocabulary is to your English ability?

1	2	3	4	5
Not at all	Not Much	Quite	Very	Extremely

A. Why?

3. A. How motivated are you to study vocabulary?

1	2	3	4	5
Not at all	A Little	Moderate Amount	Very	Extremely

A. Why?

4. **What do you think about the three word card methods (pre-made, self-made with information provided, or self-made with free choice of information)?**

5. **Which of the word card methods do you prefer? Why?**

6. **If you had free choice of the vocabulary study method for this course, what would you choose and why?**

7. **Which do/would you prefer?**

   A. Make my own word cards
   B. Study from Premade Cards

Vocabulary Tests

1. **What do you think of the weekly spoken vocabulary test system?**

2. **What do you think of the weekly vocab test for helping you learn the words, or assess your knowledge of the vocabulary?**

3. **What do you think of the written vocabulary test system?**

4. **What do you think of the written test for helping you learn the words, or assess your knowledge of the vocabulary?**

5. **The vocabulary tests: (Can choose as many as you want, or none of them).**

- improve my knowledge of my words
- help me to understand what I know and don't know
- make me study
- are not a good use of class time
- should not be used when calculating my final grade
- are an accurate measure of my vocabulary knowledge
- are not useful
- are enjoyable (weekly spoken test)
- are enjoyable (written test)
- Other (please specify)