台灣國中生合併關鍵字記憶法與間隔回憶練習學習英語字彙之成效

The effects of combining the Keyword mnemonic with Spaced Recall Practice on English vocabulary learning among Taiwanese junior high school students

by

傅查德 Chad Frederick Hawkins

Thesis

Presented to the Faculty of the

Department of Foreign Languages and Literature of

Tunghai University

in Partial Fulfilment of the Requirements for the Degree of

MASTER OF ARTS in

Teaching English As A Foreign Language

TUNGHAI UNIVERSITY

June 2016

中 華 民 國 一 零 五 年 六 月

TABLE OF CONTENTS

AKNOWLEDGEMENTS

Completing this thesis was an arduous task. I would like to express my sincerest gratitude to all the people who made it easier with their help, patience, and understanding of many people.

First and foremost, I would like to thank my thesis advisor, Dr. Chia-hui Chiu. Her patience with this student who did not always keep his advisor informed of what he was doing was boundless. Furthermore, she always kept track of the changes I was making to my thesis, and gave me invaluable help in changing and editing my thesis so that it is more understandable to the reader. Without her help I would not have been able to complete the thesis. I would like to express my deepest gratitude and admiration for her help, advice, and above all, patience.

Next, I would like to thank my committee members, Dr. Min-Hsun Chiang of the Department of Foreign Languages and Literature at Tunghai University and Dr. Hsin-I Chen of the Department of Applied Foreign Languages at National Kaohsiung University of Applied Sciences. They both gave me very helpful advice on restructuring my thesis and adding clarifying and missing sections. I appreciate them for the advice they gave me.

I would also like to extend my gratitude to Kelly for always keeping me informed about the administrative aspects of my thesis. Without her help I would have

been lost when it comes to the university administration.

Finally, I would like to express my thanks and love to my wife that stood by me and supported me through the long years I was doing my thesis. She is the one that made this all worthwhile for me.

The Effects of Combining the Keyword Mnemonic with Spaced Recall Practice on English Vocabulary among Taiwanese Junior High School Students

Chad Frederick Hawkins

Advisor: Chia-Hui Chiu

ABSTRACT

The aim of this study was to investigate the effects of combining the Keyword Mnemonic with Spaced Recall Practice in a real world classroom setting. More specifically, this study aimed to examine if combining the Keyword Mnemonic with Spaced Recall Practice (forthwith called the Combined Method) was significantly better than only using Spaced Recall Practice alone. Furthermore, the study attempted to uncover any differences in the rate of forgetting between the Combined Method and Spaced Recall Practice over the period of a month.

This study recruited 42 grade eight junior-high students to participate in the instructional treatments. Prior to the beginning of the experiment, the researcher-instructor taught them how to use the Keyword Mnemonic. Then, once per week over a month between March and April 2016, they used the self-designed vocabulary workbooks to memorize seven words using the combined method and

seven words using Spaced Recall Practice. Finally, directly after the treatment, a third of the participants wrote an immediate posttest on the fourteen words they had memorized and an additional seven words that they hadn't encountered before. These additional words acted as a control. One week later, another third of the participants underwent the same test. Finally, four weeks later, the rest of the participants underwent the same test.

Three repeated measures ANOVA tests were performed to compare the learning of words in the Combined Method, Spaced Recall Practice, and Control condition for the three posttests respectively, followed up with Post Hoc analyses. The results implied that there was no significant difference between the Combined Method condition and the Spaced Recall Practice condition in both the immediate posttest and the first delayed posttest. However, there was a significant difference between the two learning conditions in the second delayed posttest ($n=17$; $p=0.031$). The participants also scored significantly lower on the Control condition than they did for both the Combined Method and Spaced Recall Practice conditions in the immediate, first delayed and second delayed posttests. The current study also found that the data points of the Combined Method and Spaced Recall Practice both fit power function curves. However, the curve of the Combined Method was shallower than that of the Spaced Recall Practice condition, suggesting that using the Combined Method

significantly benefitted participants' retention of the target L2 English words compared to using Spaced Recall Practice alone. Based on the results, pedagogical implications and limitations of the study are provided.

Keywords: keyword mnemonic, spaced recall, English vocabulary learning

台灣國中生合併關鍵字記憶法與間隔回憶練習學習英語字彙之成效

研究生：傅查德

指導教授：邱嘉慧

摘要

本研究旨在探討合併關鍵字記憶法與間隔回憶練習於實際課室字彙教學之成效。具體地說，本研究檢視關鍵字記憶法與間隔檢索練習的合併（即合併方式）是否比單獨使用間隔回憶練習的效果更好。此外，本研究亦檢視教學實驗結束一個月後，合併方式法與間隔檢索練習在遺忘率上是否有顯著差異。

本研究招募 42 位國中二年級學生參與教學實驗。實驗開始前，研究者以授課者身分教導學生如何使用關鍵字記憶法。接著，在二零一六年三月至四月當中的四週期間，學生每週一次使用研究者自行設計之字彙練習簿在合併方式情境下學習記住七個字彙，在間隔檢索練習情境下學習記住七個字彙。在字彙學習活動後，三分之一的參與者接受測驗，測驗內容包括十四個學習記憶的字彙以及七個不曾學習過的字彙。這七個不曾學習過的字彙乃是控制對照字彙。一週後，另三分之一的參與者接受相同的測驗。最後，在四週後，其餘的三分之一參與者接受相同的測驗。

研究者以重複量數變異數分析（Repeated-measures ANOVA)以及事後檢定法進行統計分析，比較以合併方式法、間隔回憶練習法，以及控制情境下單字學

習之成效。分析結果顯示，參與者在合併方式法和間隔回憶練習法的學習成效在立即後測和第一次延遲後測沒有顯著的差異。然而，在第二次延遲後測，合併方式法和間隔回憶練習法則有顯著差異（n =17; P = 0.031）。參與者在控制字彙的學習成效無論在立即後測、第一次延遲後測、或第二次延遲後測都比在合併方式法和間隔回憶練習法的表現顯著較差。此外，合併方法及間隔回憶練習法的數據點都符合函數曲線，且合併方式法的曲線比間隔回憶練習的曲線還要低淺。亦即是合併方式法在英語作為第二語言的字彙記憶存留上較單獨使用間隔回憶練習來得有效。研究者依據結果提出教學意涵與研究限制。

關鍵字: 關鍵字記憶法、間隔回憶、英語字彙學習

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

The current study explored the effects of combining the Keyword Mnemonic with Spaced Recall Practice on the retention of English vocabulary words by Junior high school students in Taiwan. Three learning methods which were reviewed by Dunlosky, Rawson, Marsh, Nathan, and Willingham, (2013) that have been demonstrated to be effective in the learning of vocabulary words are distributed practice, retrieval practice, and the Keyword Mnemonic. The combination of distributed practice and retrieval practice has been well studied and shown to be particularly effective for the memorization of vocabulary (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke, & Bauernschmidt, 2011; Karpicke, & Roediger, 2010; Larsen, Butler, & Roediger, 2009; Logan & Balota, 2008). This combination is called Spaced Recall Practice (Logan & Balota, 2008). However, when looking at the techniques used for memorization when using these methods, it is usually just rote memorization. One memorization technique specifically designed for vocabulary acquisition that has been shown to be more effective under certain circumstances than rote memorization for compatible words is the Keyword Mnemonic (Avila & Sadoski, 1996). It follows logically that if the Keyword Mnemonic is better than rote memorization, a combination of Spaced Recall Practice and the Keyword Mnemonic

(the Combined Method) might yield better results than when rote memorization is used with Spaced Recall Practice. However, this assumption cannot be taken at face value. While many studies have found positive results for the Keyword Mnemonic (Raugh & Atkinson 1975; Atkinson & Raugh 1975; Avila & Sadoski, 1996) others have found it to be less effective than other methods (Wang, Tomas, Inzana, & Primicerio, 1993; Campos, Gonzalez & Amor, 2003; Fritz , Morris, Acton, Voelkel, & Etkind 2007). Furthermore, even though Avila and Sadoski, (1996) found the Keyword Mnemonic to be more effective than rote memorization, Campos, Gonzalez and Amor, (2003) found the opposite. Despite these findings, encouraging results with combinations of the Keyword Mnemonic plus other methods have been found (Brown & Perry, 1991; Rodrigues & Sadoski, 2002). However, when it comes to the Combined Mehtod, the scant literature that exists about this combination is contradictory, and, more worryingly, hasn't properly investigated the Combined Method's long-term (Fritz et al., 2007). Thus, there exists a gap in the literature regarding the effectiveness of combining Spaced Recall Practice with the keyword mnemonic over an extended period of time, especially when it comes to retention intervals (RI) of longer than a week (Fritz et al., 2007). Consequently, the purpose of this study was to ascertain if there are any benefits to combining the keyword mnemonic with Spaced Recall Practice when applied to the Taiwanese junior high

school English learning context. Finally, the results of this study contributed to a better understanding of the effects of combining the Keyword Mnemonic with Spaced Recall Practice and it could inform both researchers and educators regarding its relative efficacy for long-term retention of foreign vocabulary words.

In this chapter, the background of the study will first be introduced followed by a look at the gap in research regarding the above mentioned mnemonic techniques. Next, the purpose of the study and the research questions will be introduced, after which the significance of the study will be explained. Finally, a definition of the terms used in this paper.

1.1 Background of the study

It is essential for students to build up a large enough vocabulary to be able to use a language effectively because vocabulary size is important for language proficiency, reading comprehension, writing success, and language test performance (Alderson, 2005; Laufer, 1997; Laufer and Goldstein, 2004; Nation, 2006; Schmitt, Jiang, & Grabe, 2011). In order to achieve deeper comprehension of complex texts such as academic texts, 98% of the running words need to be comprehensible (Schmitt et al., 2011). Consequently, 8000 to 9000 word families are needed to understand written texts at 98% coverage and at least 6000 to 7000 word families are needed to understand 98% of spoken language (Nation, 2006). Native college

educated English speakers know anywhere between 12 000 (Zechmeister et al., 1995) and 17 000 word families (Gould, Nation, & Read, 1990). Although these amounts of words are above what most non-native English speakers achieve (Nation, 2001; Nation & Waring in Schmitt, N. & McCarthy, M., editors, 1997), the 2000 most frequently used words in English cover about 78.1% of academic texts and up to 90.3% of conversational English (Nation, 2001). These numbers are doable, thus Nation (2006a) suggests that the first 2000 or so high frequency words require class time because of their wide occurrence and high frequency.

Although learning words in context may be beneficial, especially when complimented with hypertext or isolated glosses (Yun, 2011; Hill & Laufer, 2003), Laufer (2005) showed that guessing from context could be problematic (Laufer & Yano, 2001; Laufer, 1997b) because there may not be enough context clues, the clues themselves may be misleading, and the readers schemata may not be compatible with the text context. However, deliberate learning of words from lists, although unfashionable, can be very effective for the acquisition of large amounts of vocabulary in relatively short periods of time (Nation, 1983; Thorndike, 1908). Research has shown three methods which can help with vocabulary acquisition via lists are distributed practice (Bahrick & Phelps, 1987; Bloom & Shuell, 1981; Dempster, F.N. ,1987; Kang, et al., 2014; Logan & Balota 2008), retrieval practice

(Bjork & Bjork, 1992; Carrier & Pashler, 1992; Roediger & Karpicke 2006a; Roediger & Karpicke 2006b), and the Keyword Mnemonic (Raugh, Schupbach, & Atkinson, 1977; Raugh & Atkinson 1975; Atkinson & Raugh 1975; Atkinson,1975).

Of the three methods mentioned above, distributed practice may arguably be the most studied with the first empirical study being conducted 130 years ago (Ebbinghaus, 1885; Bahrick & Hall, 2005; Dunlosky et al., 2013). Distributed practice refers to the act when distributing learning instances over time instead of massing them one after the other leads to an improvement in the long-term retention of the learned material. It has been shown to be effective for a wide range of learning situations (Dunlosky et al., 2013), especially with the learning of foreign language vocabulary (H. P. Bahrick, 1979; Bloom & Shuell, 1981; Bahrick & Phelps, 1987; Dempster, 1987; Kang et al., 2014).

Retrieval practice takes advantage of the fact that recalling information that has been learned increases the likelihood of long-term retention relative to studying the information again (Dunlosky et al 2013; Roediger & Butler, 2011).This effect is called the testing effect. As early as 1909, empirical studies showed the benefit of Retrieval Practice in an experimental setting (Abbott, 1909) and were shown to be effective in the educational setting 30 years later (Spitzer, 1939). Not only does Retrieval Practice seem to be better than restudying for long-term retention (Roediger, & Karpicke,

2006), but it is also better than elaborative learning via concept mapping (Karpicke, & Blunt, 2011). Furthermore, Retrieval Practice has also been shown to be effective in the learning of foreign language vocabulary words in written form (Carrier, M., and Pashler , H.,1992; Pashler, Cepeda, Wixted, & Rohrer, 2005), aural comprehension, and ability to produce the L2 words (Kang, Gollan, & Pashler, 2013). However, without feedback, incorrect information or ideas could be fixed into longer term memory (Spitzer, 1938). Corrective feedback enhances not only the correct retention of incorrectly recalled information (Pashler, Cepeda, Wixted, & Rohrer, 2005), but also the retention of low confidence correctly recalled information (Butler, Karpicke, and Roediger, 2008). Furthermore, Retrieval Practice has been successfully combined with distributed practice many times (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke, & Bauernschmidt, 2011; Karpicke, & Roediger, 2010; Logan & Balota, 2008), and this combination, called Spaced Recall Practice, has been shown to be better than just repeatedly studying (Larsen, Butler, & Roediger, 2009; Butler, 2010).

However, when it comes to incorporating the Keyword Mnemonic with Spaced Recall Practice, the literature is less clear. The Keyword Mnemonic is a Mnemonic technique pioneered by Atkinson & Raugh (1975). It involves finding an L1 or a known L2 word that sounds similar to the L2 word that is being studied and using it as a bridge to the L1 definition or equivalent word by creating a mental image of the

two words interacting (Nation, 2001). Although the original series of experiments run by Atkinson and Raugh were very encouraging (Raugh, Schupbach, & Atkinson, 1977; Raugh & Atkinson 1975; Atkinson & Raugh 1975; Atkinson, 1975), later studies suggested that there may be several limitations to the Keyword Mnemonic, including its limitation to easily imaginable words, the fact that the efficacy of self-generated keywords versus teacher supplied keywords is still unresolved, and that it may not help with long-term retention (Wang, Thomas, Inzana, Primicerio, 1993; Dunlosky et al., 2013). Interestingly, studies done by Brown & Perry, (1991) and Rodrigues & Sadoski, (2002) showed that even though, under the conditions present in those studies, using the Keyword Mnemonic on its own was the least effective way of memorizing vocabulary words, using the combination of the Keyword Mnemonic with either semantic processing or with contextual guessing proved to be the most effective ways of memorizing the vocabulary in the respective studies. However, specific to this study, the combination of the Keyword Mnemonic and Spaced Recall Practice has not been proven to be more effective for long-term retention than Spaced Recall Practice alone in all conditions (Fritz et al,. 2007). However, Fritz et al., (2007) did hint at possible benefits to recall of L2 English words in their immediate posttest that may be extended to long-term retention if the intervention period is longer. This is what inspired the current study.

1.2 Statement of the problem

As stated above, while there is a large literature supporting the efficacy of Spaced Recall Practice (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke, & Bauernschmidt, 2011; Karpicke, & Roediger, 2010; Logan & Balota, 2008), the empirical evidence relating to the keyword Mnemonic is less clear. While some studies show clear benefits relating to the keyword mnemonic (Raugh & Atkinson 1975; Atkinson & Raugh 1975; Avila & Sadoski, 1996) others have had less encouraging results (Wang, Tomas, Inzana, & Primicerio, 1993; Campos, Gonzalez & Amor, 2003; Fritz et al 2007). Furthermore, even though it has been show that combining the keyword method with semantic processing (Brown & Perry, 1991), and with contextual guessing (Rodrigues & Sadoski, 2002) seemed to be beneficial, when it comes to combining Spaced Recall Practice with the Keyword Mnemonic, the only study that could be found that looked at it (Fritz et al 2007) had contradictory results. Fritz et al., (2007) found that combining the Combined Method only had significant benefits for receptive vocabulary in the delayed test compared with Spaced Recall Practice alone. Furthermore, although there was a significant initial benefit to the productive performance associated with the Combined Method, it was consequently lost by the one week delayed posttest.

Unfortunately this study didn’t investigated what the delayed effects beyond a

week of the Combined Method would be on either the productive or receptive vocabulary retention. This is an important omission because without a long delayed test the efficacy of this treatment on actual learning and long-term retention of vocabulary remains unknown. Finally, the treatments took place in one session. This makes the results of the experiment less applicable to the classroom reality. As already mentioned, repeated recall spaced over periods of a week or more could significantly alter the effects of a treatment (Bahrick, 1979, Cepeda et al, 2008). Thus it is imperative to investigate the effect of the treatments if they had taken place in a more realistic timeframe extended over a few weeks or a month.

In sum, there existed a gap in the literature regarding the efficacy of the combination of Spaced Recall Practice and the Keyword Method for the long term retention (realistic periods of more than a week) of foreign vocabulary words. Furthermore, incorporating such a combined method into a real-world classroom setting over a realistic timeframe hasn't been attempted either.

1.3 Purpose of the study and research questions

The general purpose of this study was to examine the effects of combining the Keyword Mnemonic with Spaced Recall Practice on the long-term retention of foreign vocabulary words. More specifically, the study aimed to establish if adding

the Keyword Mnemonic to Spaced Recall Practice would result in superior long-term retention (at least four weeks) of vocabulary words than just using Spaced Recall Practice alone. The dependent variable in this study was the amount of foreign language English vocabulary retained. The independent variables were the learning strategies involved, and the retention intervals. The instrument used to measure the dependent variable was an L1 Chinese to L2 English cued recall test which measured the productive vocabulary retention. This instrument is called the Vocabulary Retention Test.

From this purpose statement, the following research questions can be distilled:

1. Do the participants retain significantly more L2 English words on the immediate posttest when using the Combined Method than when using Spaced Recall Practice?

2. Do the participants retain significantly more L2 English words on the first delayed posttest when using the Combined Method than when using Spaced Recall Practice?

3. Do the participants retain significantly more L2 English words on the second delayed posttest when using the Combined Method than when using Spaced Recall Practice?

4. Are there any significant differences between the immediate posttests, the one

week delayed posttests, and the two week delayed posttests?

1.4 Significance of the study

This study is significant to learners and educators because it can help lessen the load of learning a foreign language by enlightening potential learners and educators about the most efficient vocabulary memorizing methods. As already mentioned, becoming fluent in a language requires quite a large vocabulary (Nation 2006) and most foreign language learners never acquire native like vocabularies (Nation & Waring in Schmitt & McCarthy, editors, 1997). If there is a significant advantage to combining the Keyword Mnemonic with Spaced Recall Practice, this study could enlighten educators on how to implement the Combined Method in classroom settings. Alternatively, if there isn't any advantage to using the Combined Method, educators could focus on Spaced Recall Practice as a basis for memorizing vocabulary and only incorporate the Keyword Mnemonic into it for hard to learn suitable words.

This study is significant to the EFL field because it builds on information in the literature about the long-term effectiveness of the keyword method. It demonstrates that extending the treatment period of the combined method significantly alters the results obtained by Fritz et al., (2007), thus demonstrating the potential benefits of adding the keyword method to Spaced Recall Practice. Furthermore, it adds to a growing mass of literature that supports the assertion that the Keyword Mnemonic has

an enhancing effect on other methods if combined with them.

1.5 Definition of the terms

1. *Distributed practice* (DP): DP is an umbrella term referring to the practice of spacing learning instances over time thus benefitting from the *spacing effect* and the *lag effect* (Cepeda et al., 2006). When retention of the information learned is better after repeated learning of the same material takes place with temporal gaps between study sessions than when the material is repeatedly studied with no gaps between the relearning sessions, the *spacing effect* is observed (Dempster, 1988). The *lag effect* refers to how different numbers of intervening items, or different lengths of time between presentations, effect the *spacing effect* (Cepeda et al., 2006; Kraft & Jenkins, 1981).
2. *The testing effect*: This effect refers to the finding that taking a test on information that has been learned is better for retention than restudying the material for the same amount of time (Roediger & Karpicke; 2006b).
3. *Retrieval Practice:* The actual act of recalling the information from memory is called *retrieval practice*, and the additional effort involved in recalling from memory may be why it is so effective (Roediger & Butler, 2011).
4. *The keyword mnemonic*: This mnemonic technique uses mental imagery to form links between an unknown L2 word, its meaning, and a similar sounding L1 word

(Atkinson, 1975). For example, when learning the Indonesian word for door, *pintu*, one could imagine a door with a *pin* in it (Nation 2001), or a "pin" moving "to" a door. Thus, when trying to recall door in Indonesian, this image may help recall the phonetic sound of "*pintu*".

5. *Spaced Recall Practice*: This refers to the observation that when combining the *distributed practice* with *retrieval practice*, both effects are enhanced (Kang, Lindsey, Mozer, & Pashler, 2014).
6. *Retention interval* (RI): *Retention interval* refers to the time between the last study or learning session in an experiment and a subsequent test of retention (Cepeda et al., 2006). For example, if there is an immediate posttest two minutes after an intervention, the retention interval is two minutes. However, if there were to be a delayed posttest ten minutes later, the retention interval would be ten minutes, and not twelve minutes because the last study session was the immediate post test
7. *Productive vocabulary*: Vocabulary that can be retrieved from memory and produced to convey a meaning is essentially *productive vocabulary* (Nation, 2001). For the purposes of this study, *productive vocabulary* will be defined as L2 English words that can be recalled from memory if an L1 equivalent is given as a prompt.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, the theoretical background and justification for the design of this study will first be explored. This will include three subsections: an overview of Jiang's psycholinguistic model of vocabulary acquisition in a foreign language (Jiang, 2000); theories and models trying to explain the phenomena associated with Spaced Recall Practice; and Dual Coding theory (Clark and Paivio, 1991) as an explanation for the Keyword Mnemonic's effects on retrieving words. Next, the importance of vocabulary will be explored and experiments showing how vocabulary size relates to the four language skills (listening, speaking, reading and writing) will be examined. After that, a thorough review of empirical studies on the effects of Spaced Recall Practice, the keyword mnemonic, and the combination of the two will be provided. Finally, this chapter will conclude with a brief explanation of why the theory justifies the current experiment.

2.1 Theoretical background of the study

Theoretically, how do learners acquire new L2 words and what is the best way to acquire them? In this section an attempt will be made to answer these questions as best as possible according to current theories.

2.1.1 Jiang's psycholinguistic model of vocabulary acquisition

To understand how a word is learned, it is important to know what needs to be learned to truly know a word. Nation, (2001) points out that there are various aspects of knowing a word (see table 2.1). First, to know a word one needs to know its form. Knowledge of this form includes the spoken form (the phonological aspect) of the word, the written form of the word (the orthographic aspect), and knowledge of the word's subparts. Not all these parts need to be known, but at least the spoken or written form needs to be known. Second, to be understood, a words' meaning needs to be known. That is, its form (either spoken or written) needs to refer back to a known concept or idea. This concept or idea (whether referring to an object, action or feeling) may have various referents and associations connected to it. Not all of these referents or associations need to be known in order for a word to retain some meaning. Finally, how a word is used in order to convey a meaning is critical not only for understanding but also to convey meaning. For example, is a word used as a verb such *vanquish*, stating agency, or is it a noun, refereeing to a concept such as *victory*. Furthermore, which other words are likely to be used with this word. Knowing if *occur or happen* collocates more often with earthquake is an example of this. Also, knowing when it is appropriate to use a word is yet another aspect of knowing the use of a word.

Table 2.1. *Aspects of word knowledge*

aspect of word	aspect of knowledge
Form	spoken, written, word parts
Meaning	form and meaning, concept and referents, associations
Use	grammar functions, collocations, constraints of use

Adapted from *Learning Vocabulary in Another Language,* p.27, by I.S.P. Nation, 2001, Cambridge: Cambridge University Press.

As seen above, there are many aspects of knowing a word, and learning a word is not a binary occurrence. Learning a word is more of a gradual process (Nation, 2001) that could be interrupted at any stage and doesn't need to be completed for the word to have agency in the mind of its user. The psycholinguistic model of second language vocabulary acquisition proposed by Jiang, (2000; 2002) describes how this gradual process might take place.

Jiang's psycholinguistic model of L2 vocabulary acquisition is based on Levet's (1989) model of lexical retrieval (Jiang, 2000; Jiang, 2002). Jiang describes four aspects of knowing an L1 word namely, semantics, syntax, morphology, and form (phonology and orthography). The semantics and syntax are said to be part of the lemma level of the word whereas the morphology and form are part of the lexeme level of the word according to Jiang. Wherever these different elements of word knowledge are stored on the neurological level in the brain, they are linked to each

other creating the meaning entity we perceive as an understood word. For example, when we see the word *dog* in its orthographic form, we immediately also access the semantic aspects of the word (a mammal which wags its tail and barks) and also its phonological aspects (how we think the word sounds). In actual fact, all the lemma and lexeme level knowledge gets activated in short order when any part gets either receptively or productively stimulated (Jiang, 2000; Jiang 2002).

Jiang suggests when a new L2 word is learned the main focus is at first on the formal representation of the words (phonology and orthography); the semantic and syntactic categories remain empty. Although both semantic and syntactic information on the L2 word can be available, it is usually through an L1 equivalent or explicitly learned information. The word is not integrated into the lexical or syntactical representations and must be accessed consciously. Thus, the meaning of the word is mediated through its L1 equivalent. Jiang calls this first stage the formal stage. As an L2 word is repeatedly activated, it forms stronger links to the L1 equivalent word form as well as the lemma level knowledge of the L1 equivalent word. Thus, over time, a direct link is established between the L1 lemma level entries and the L2 word. This second level of L2 acquisition is where a lot of L2 vocabulary is fossilized. These words have in essence just copied the lemma level word knowledge from their L1 equivalents. The second stage is called the L1 lemma mediation stage (Jiang 2000).

As more exposure to the L2 word occurs and it is repeatedly activated, more lexical, syntactic, and morphological information about the L2 word may be integrated into the lexical representations of the L2 word. This is the lexical integration stage (Jiang, 2000) and the final stage in Jiang's L2 vocabulary acquisition model.

According to Jiang's model, when we memorize word pairs, we are copying the lemma level lexical information of the L1 word equivalents into the L2 lexical store. Although this does not produce perfectly accurate understanding of the words in the beginning, it does save a lot of time (Jiang, 2002). Once the L1 lemma mediation stage is reached, exposure to authentic L2 use of the target vocabulary is most probably the best way to form a deeper understanding of the word. However, before that stage, both the Keyword Mnemonic and Spaced Recall Practice should be able to help with the formation of close links between the L2 target word and the L1 lemma information. More will be discussed on why the Keyword Mnemonic could help a word reach the L1 lemma mediation stage in section 2.1.3. A discussion about why Spaced Recall Practice should help can be found in section 2.1.2.

2.1.2 Theoretical background for Spaced Recall Practice.

Spaced Recall Practice consists of a combination of Retrieval Practice, which takes advantage of the testing effect, and Distributed Practice, which takes advantage

of the spacing effect. In this section, Retrieval Practice will first be discussed. Distributed Practice will then briefly be introduced and its natural combination with Retrieval Practice will be explained.

Although the benefits of retrieval practice have been well established, no single satisfying theory about why it works has been well established (Dunlosky et al., 2013). However, one proposed mechanism for the testing effect comes from Carpenter (2009). It states that retention is enhanced by an elaborative retrieval process. In the process of searching for target information, related information gets activated. This information may then be encoded or connected to the target information, thus increasing the amounts of connections to the target information. This in turn makes future retrieval of the target information easier.

However, if this theory is to be accepted, it increases the importance of corrective feedback enormously. This is because without corrective feedback, errors could be encoded into longer term memory (Spitzer, 1938). If an error is made in the retrieval process, it might be encoded with the original information, thus making future incorrect retrieval possible (Carpenter, 2009). However, as Pashler, Cepeda, Wixted, & Rohrer, (2005) demonstrated, corrective feedback enhances the correct retention of incorrectly recalled information. Furthermore, it also decreases the likelihood that low confidence correctly recalled information would be incorrectly

recalled later on (Butler, Karpicke, and Roediger, 2008).

Next, Distributed Practice has been shown to be effective for a wide range of learning situations (Dunlosky et al., 2013) and it's especially effective with the learning of foreign language vocabulary (H. P. Bahrick, 1979; Bloom & Shuell, 1981; Bahrick & Phelps, 1987; Dempster, 1987; Kang et al., 2014). However, exactly how Distributed Practice works is up for debate. However, Dunloskey et al., (2013) identify some hypotheses that are contenders for a theoretical explanation. First, there is the theory of deficient processing (Bahrick & Hall, 2005). This theory proposes that additional learning sessions' efficacy suffers if spaced to close to previous learning sessions. Another theory that is gaining traction states that the additional exposures serve as reminders of the original context. These reminders stimulate the recall of the original experience, thus inferring the benefits of the testing effect on the memory trace (Dunloskey et al., 2013; Benjamin, & Tullis, 2010). If the reminder is too early it isn't as significant; its impact is too small. Conversely, if the reminder is too late, it is unlikely to reactivate the original experience (Benjamin, & Tullis, 2010).

The combination of Retrieval Practice, and Distributed Practice, called Spaced Recall Practice yields superior results than either one alone (Roediger & Butler 2011; Roediger & Karpicke 2006b). The theory of reminding of Benjamin, & Tullis, (2010), and the theory of elaborative retrieval Carpenter, (2009) proposed seem to

complement each other. Because each distributed practice elicits retrieval of the original learning experience, retrieval has occurred. Because retrieval has occurred, elaborative retrieval may cause additional memory traces to be encoded into the original experience, thus making future retrieval easier. In other words, after each successive retrieval, the memory trace is shored with additional associations, making further retrieval relatively easier.

2.1.3 Dual coding theory

Dual coding theory is basically a theory of cognition that states that memories or experiences have both verbal and non-verbal (imagery) representations and that these representations have their own systems (Clark & Paivio, 1991). In other words, Paivio postulates that there is a verbal system and an imagery system in the mind. Images of shapes (a ball), sounds (train whistle) visual actions (opening a door) etc., are all non-verbal or imagery representations (Paivio, 1969). These representations are connected to real word things, actions, or events and are not just arbitrary codes or symbols. That is, a mental representation of *telephone* brings to mind similar visual and other sensory qualities to the real world object it refers to (Clark & Paivio, 1991). However, verbal representations contain aural (spoken words) and visual (written words etc.) codes that are arbitrary representations of their real referents. That is to say, the verbal representations arbitrarily symbolize the real world objects, actions or

events they refer too. The word Banana in English or 香蕉 in Chinese refer to the same sweet fruit (Paivio 1969; Clark & Paivio 1991). The fact that there are so many languages and symbols referring to the same things shows the arbitrary nature the verbal system.

The benefits to memory associated with the Keyword Mnemonic could be explained in terms of Dual Coding Theory (Paivio 1969). Two processes that Dual Coding Theory uses to explain these benefits are elaboration and organization. First, with elaboration both the imagery and verbal codes add together to produce an effect that is better than that of just the verbal code (Clark & Paivio, 1991). As Paivio and Lambert (1981) demonstrated, generating images of words lead to better memory retrieval than simply repeating those words. Second, imagery codes also benefit recall from memory because they can represent separate elements in a unified or interactive image. Thus, if one of the separate elements is activated, the whole image gets reactivated. Then, in turn, any verbal codes associated with the imagery code also get reactivated (Clark & Paivio, 1991).

To sum up, Dual Coding Theory proposes that there are two memory systems, a verbal and an imagery system, that store and create representations of objects, actions or events. Also, Dual Coding Theory could be used to explain why the Keyword Mnemonic benefits memory. In short, because of the elaborative effect (having an

image associated with a verbal code) and organizational effect (an image is worth a thousand words), Dual Coding Theory suggests that the imagery link in the Keyword Mnemonic will benefit the retrieval of associated L1 and L2 words.

2.1.4 Theoretical justification for the study

The theoretical justification for this study looks at why combining the Keyword Mnemonic with Spaced Recall Practice should at least have an additive effect. First, looking at the Keyword Mnemonic from the perspective of form-meaning mapping proposed by Jiang (2002), and Dual Coding Theory (Clark & Paivio, 1991), the following becomes apparent. First, when a new word is learned using the Keyword Mnemonic, links are made to lemma level lexical entries from both the keyword and the L1 equivalent word. Thus, a clue to the phonology of the target L2 word is linked to, or encoded into the memory trace of the target L2 word. At the same time, a link is made to the semantic component of the lexical entry of the L1 equivalent word. The links are made by associating an interactive image with both the Keyword lexical entry and the L1 lexical entry. To elaborate, by incorporating non-verbal (imagery) representations (Clark & Paivio, 1991) from both the Keyword lexical entry and L1 equivalent word lexical entry (Jiang, 2000) into an interactive image, links are made from this image via the non-verbal representations to the rest of the associated verbal and non-verbal specifications in both the lexical entries. Although this new lexical

web (the connections between the keyword, L1 word, L2 word, and interactive image) can be activated by recalling one of its elements, this complex memory trace (lexical web) is susceptible to forgetting (Dunlosky et al., 2013).

With the addition of the theory of reminding, Benjamin, & Tullis (2010), and the theory of elaborative retrieval, Carpenter (2009) the following can be postulated. When using Spaced Recall Practice with the Keyword Mnemonic, the complex memory trace created using the Keyword Mnemonic should be strengthened by the repeated encoding of associations every time an element of the complex memory trace is retrieved. Thus, combining the Keyword Mnemonic with Spaced Recall Practice into a Combined Method should be beneficial because Spacer Retrieval Practice may shore up the complex memory trace created with the Keyword Mnemonic against forgetting.

2.2 The importance of vocabulary

It has already been well established that building up a large vocabulary is crucial in the acquisition of a foreign language because vocabulary size is important for language proficiency, reading comprehension, writing success, and language test performance (Folse, 2004; Laufer, 1987; Laufer and Goldstein, 2004; Nation, 2001; Nation, 2006; Read, 2000; Schmitt, 2010; Schmitt, Jiang, & Grabe, 2011).

2.2.1 Vocabulary size and reading comprehension

There are various reasons for the importance of vocabulary. To start with, empirical studies have supported the assertion that there is a strong relationship between vocabulary size and reading, writing, grammar, and listening. Four studies that support this assertion are those of Laufer (1992), Hu & Nation (2000), Schmitt, Jiang, & Grabe (2011), and Alderson (2005).

First, in a study of 92 first-year university students in an English for academic purposes class, Laufer (1992) investigated whether vocabulary size correlates to reading comprehension. This was done by comparing how well the students did on either the Vocabulary Levels Test (Nation, 1989) or the Eurocentres Vocabulary Test (Meara, 1989) with the students' scores on two standardized reading comprehension tests. Five groups' reading scores were compared, those with a vocabulary level of below 2000 word families, those with a vocabulary level of 2000 word families, those with a vocabulary level of 3000 word families, those with a vocabulary level of 4000 word families, and those with a vocabulary level of 5000 word families. A correlation of $r=0.5$ ($p < 0.0001$) was found when comparing the scores from the reading comprehension tests and the Vocabulary Levels Test and a highly significant correlation was found when the comparison was made with the Eurocentres Vocabulary Test ($r=0.75$, $p <0.0001$). When comparing the results of the five groups

with each other, the most significant difference between the groups appeared between the 2000 and the 3000 word family levels. The results of this study showed that there was a correlation between vocabulary size and reading comprehension, although, as Laufer stated, this correlation does not indicate causation.

Second, Hu & Nation (2000) studied the reading comprehension levels of sixty-six participants from a wide linguistic background with a high English proficiency, attending a pre-university English course at an English University in an English speaking country. The focus was to see if different word densities would result in different levels of comprehension. The students were asked to read different versions of a fiction, which had either 0%, 5%, 10%, or 20% nonsense words in the text (in other words, either 100%, 95%, 90%, or 80% of the vocabulary words were known by the readers). They were then given both a multiple-choice questionnaire (with a maximum possible score of 14) and a written recall test (with a maximum possible score of 124). The results for the multiple-choice questionnaire showed that 48.62% of the variance could be attributed to the density of unknown words ($F=58.75$ with 1 and 64 degrees of freedom, $p<0.0000$). The results for the written recall test showed that 62.18% of the variance could be attributed to the density of the unknown words ($F=105.31$ with 1 and 64 degrees of freedom, $p<0.0000$). These results indicated that the learners' comprehension of the text increased predictably with an

increase in the known words in the text.

Third, in a study of 661 participants from 8 countries Schmitt, Jiang, and Grabe (2011) also compared reading comprehension with the amount of known words in a text. In order to do this, they selected two texts that were of equivalent difficulty according to Flesch–Kincaid Grade Level. They tested the students' vocabulary knowledge using a checklist of 150 words, of which 30 were non-words. Any participant who selected three or more non-words was eliminated from the study. Finally, they developed a two-part comprehension test, of which the first was a multiple-choice test and the second part was a graphic organizer completion task. Although they only found a moderate correlation (Spearman's correlation r_s=0.41) between vocabulary size and reading comprehension, Schmitt et al, speculate that the high ceiling effect (60% of the participants were in the 98%-100% vocabulary coverage range) may have lowered the correlation, and that if they had more participants in the lower vocabulary coverage ranges, the correlation might have been higher. These results suggest that although the amount of known words in a text is important, it is not the only factor that contributes to reading comprehension. That said, the amount of known words is a crucial prerequisite part of the understanding.

One of the best studies supporting the assertion that vocabulary size correlates with the four language skills mentioned above comes from Alderson's (2005) analysis

of the data gained from the DAILANG project. DAILANG is a diagnostic language assessment system which originated from a White Paper published by the European Commission named *Teaching and Learning –Towards the Learning Society*. Originally intended to be a language certification test taken online, it changed into a project to develop diagnostic language tests in 14 European languages. Delivered over the Internet, it tests five aspects of language use: reading, listening, writing, grammar, and vocabulary (Lancaster University, 2015). In version 1 of DIALANG the Vocabulary Size Placement Test was used to determine the language proficiency of the users of the system in order to give them easy, medium or difficult tests. The original pilot study contained 150 verbs of which 50 were pseudo-words. This was later reduced to 75 verbs, of which 25 were pseudo words. Users only had to state whether the words were true verbs or pseudo verbs. Six methods of scoring were devised, of which the simplest method, where users were given a point for each word correctly identified as either real or pseudo, proved to produce results that correlated the best with performance on the DAILANG tests. Of the four categories, namely reading, grammar, writing, and listening, writing had the highest correlation ($r=0.70$; $N=735$), followed by both reading and grammar ($r=0.64$; N equaling 718 and 1084 respectively) and finally listening ($r=0.61$; $N=606$). It is remarkable that a test which just tested whether users knew if the vocabulary was real or pseudo verbs correlated

so well with all the aspects of English use. This study is important because it links vocabulary knowledge to almost all aspects of language production.

When looking at all four above-mentioned studies, a clear link becomes apparent between the size of vocabulary and reading comprehension, listening comprehension, productive use of English, as well as grammatical accuracy. Although Hu and Nation's study did not directly measure vocabulary size, it did look at the amount of running words known. Thus we could assume that if a reader only knew 80% of a given text, their vocabulary size most probably is less than 2000 (closer to the 1000 BNC level) and if a reader knew 95% of the running words, their vocabulary level would be around the 3000-word level (Nation 2006). The same could be said about Schmitt, Jiang, and Grabe's 2011 study. It is important to note that increasing vocabulary does not improve reading and listening comprehension or productive use and grammatical accuracy. However, it does seem to increase the potential for improving those language skills. Therefore, it can be concluded that increasing vocabulary is a necessary step (although not the only step) in improving all round performance in English. Thus, knowing which learning methods work best for vocabulary acquisition is warranted.

2.2.2 The need for explicit vocabulary instruction

Although learning of words in context may be beneficial, especially when

complimented with hypertext or isolated glosses (Yun, 2011; Hill & Laufer, 2003), purely incidental learning of words from context is not sufficient to increase second language vocabulary efficiently (Horst, Cobb, & Meara, 1998; Zahar, Cobb, & Spada, 2001; Waring & Takkai 2003; and Rosszell, 2007). Although there are many studies supporting this assertion, only four will be touched upon in this section.

The first study is that of Horst, Cobb, and Meara (1998). In a study of 34 students in an intensive English program at Sultan Qaboos University in Oman, they investigated which factors contribute to students acquiring words from extensive reading of simplified novels. The simplified version of the novel *The Mayor of Casterbridge* was read out aloud in class over six hour long periods and the students had to follow along in books that were handed out in class and taken back at the end of each period. This was done in order to control for intentional learning of the text. Thus, it was known with a high degree of certainty that no intentional vocabulary learning took place. Forty-five words were chosen from the text based on their frequency, and two tests were developed for the experiment. One of the tests was a 45 question multiple choice test while the other was a 13 question comprehension test where the students were given a list of three words for each question and had to eliminate the odd one out. A pre-test was given about a week before the treatment commenced and a post-test was given after the conclusion of the treatment period.

The pretest showed that the students knew on average 21.6 of the target words. The post-test mean was 26.26 with a standard deviation of 6.43 with a p value smaller than 0.05, t (33) =5.81. Although the study found significant incidental increases in vocabulary, from the viewpoint of a foreign language learner, these gains were too few to be of significant use. Over the six hour period of reading the novel, the students acquired on average less than one word per hour. That is not efficient enough to be of realistic use for an EFL student. Horst, Cobb, and Meara, conclude that even when assuming an optimistic scenario of reading 50 books a year (that is one book every week), only about 250 words would be learnt incidentally. Thus, they recommend the systematic explicit teaching of high frequency words in order for the students to gain what they call “lexical independence.”

In the next study, to find out if the amount of vocabulary learnt while reading is predictable and if it is enough to establish a functional vocabulary, Zahar, Cobb, & Spada (2001) studied how many words were incidentally acquired by 144 male seventh grade students at a French speaking private school in Montreal. All the students were mainly French speakers and all had received three semesters of ESL instruction at the school. The students had previously been placed in five streams according to their level of English proficiency determined by an English placement exam developed over many years at the school. For the study, all the students were

given the written version of Nation's (1990) Vocabulary Levels Test. The results corresponded well with the levels of the streams into which the students were placed by the school. Thus the highest stream classes knew the most vocabulary and the lowest stream knew the least. A graded reader, *The Golden Fleece,* was used for this study. Thirty words were chosen as the target vocabulary according to their frequency. The students were given a pre-test on the 30 target words 16 days before the treatment began and again two days after the treatment. The treatment consisted of the students listening to a tape-recording of the story while following in their books in class. They were also given the opportunity to re-read the story afterwards. Their results showed that on average the students acquired 2.33 of the 30 words tested. Furthermore, they discovered that the frequency of the words in the text was the best indicator of whether or not a word would be learnt. Finally, they concluded that the students were learning only about one word per thousand words read. Thus, as they assert, relying purely on incidental vocabulary acquisition would be an inefficient and ineffective way of acquiring a second language for most EFL students, especially in the beginning stages.

In the third study discussed here, Waring & Takaki (2003) examined at what rate different levels of vocabulary words were retained by 15 intermediate Japanese female students reading a graded reader. Twenty-five words that appeared at five

different frequency levels (once, 4 to 5 times, 8 to 10 times, 13 to 14 times, 15 to 20 times) taken from *A little princess* were tested. The English words to be tested were changed into pseudo words in order to ensure that none of the participants were familiar with the words. Three tests were uses: a word-form recognition test; a multiple –choice test; and a meaning by translation test. The subjects were told to read the book for enjoyment, and they were told there would be a test after the reading. They were given an immediate post-test, a one week delayed-test, and a three month-delayed test. The results showed that the frequency with which words appeared influenced the rate at which the words were learned. On the immediate posttest, 18.4% of the words were correctly translated. However, after three months, only one word was correctly translated. Of the words encountered more than 15 times, only 42% were correctly translated in the immediate posttest, only 10% after two weeks, and only 3.6% after three months. No words that were encountered less than 15 times were retained after 3 months. This is a very low rate of word retention, and Waring and Takaki concluded that this translates to successfully learning only one new word per hour of effort spent.

Finally, Rosszell (2007) compared the vocabulary retention of 40 intermediate level Japanese students in two conditions. Two graded readers were used and two groups were created to counterbalance the experiment. In the experimental condition,

students read a graded reader and were assigned ten words from the reading to study each week. In the control condition, the students read the other graded reader but were only asked to write a report on the reader. The groups were given pre- post- and delayed post-tests using two different measures (the Vocabulary Knowledge Scale from Wesche & Paribakht, 1996, and a recall test modelled on Laufer and Nation's 1996 test). The experimental condition outperformed the control condition, with students gaining between 1.65 to 2.87 times more vocabulary words than in the control condition. This study clearly indicated the benefits of adding an explicit vocabulary study element to extensive reading.

In Conclusion, when looking at the above four studies, it becomes patently clear that an explicit element is needed in vocabulary instruction in order for vocabulary acquisition to become more effective. Without such an element, the acquisition rate is just too low for a learner of a foreign language. As already stated above, it would take years to acquire even just a basic vocabulary size. Seeing as most foreign language learners don't have years to study a new language and most probably have other endeavors keeping them busy as well, this slow acquisition rate is impractical. Therefore, finding out which explicit vocabulary learning practices are the most efficient and implementing them are crucially important to the construction of a successful ESL program. *Distributed Practice, Retrieval Practice,* and the *Keyword*

method are three such practices that will be further investigated below.

2.3 Spaced Recall Practice

As previously stated, Spaced Recall Practice is the combination of distributed practice and retrieval practice. Distributed practice takes advantage of the spacing effect while retrieval practice takes advantage of the testing effect. In this section, the efficacy of distributed practice will first be investigated. Then retrieval practice will be examined. Finally, this section will conclude with an overview of the combination of the two.

2.3.1. Efficacy of distributed practice

Distributed practice is the practice of distributing learning instances over longer periods of time, instead of massing learning instances all at one time. This practice leads to the well documented spacing effect, which can be observed when spaced study instances lead to better information retention than massed study instances if the amount of study time remains the same. In other words, if practice is distributed over time instead of massed, learning takes place faster relative to the total amount of time spent practicing (Baddeley, 1978; Dempster, 1989). For example, if vocabulary words are studied for three five minute periods spaced over three days, better learning should take place than if the same vocabulary words were studied for fifteen minutes in one sitting. Ebbinghaus, who also in the same study introduced the world to the

forgetting curve, first empirically studied distributed practice in 1885. At the end of Chapter 8 of *Memory: A Contribution to Experimental Psychology* (Ebbinghaus,1885; translated by Ruger & Bussenius, 1913) it is stated that when distributing practice over three days 38 repetitions were required to memorize a 12 syllable nonsense series, whereas the same effect required 68 repetitions if done in only one day. Although Ebbinghaus did these experiments on himself, they were the first empirical indication that spacing learning instances is a more efficient way of memorizing information than massing learning instances.

In 1978, Baddely & Longman showed that using distributed practice to learn a skill such as typing was more efficient with larger inter-study intervals than with smaller and massed study instances. In the experiment, postmen were taught how to type alphanumeric code on typewriters. They were assigned into four groups studying to type for two hours once a day, two hours twice a day, one hour once a day, and one hour twice a day. All groups practiced for a minimum of 60 hours. At the 60 hour mark the 1 hour once a day group was significantly faster ($p<0.01$ for length of session; $p=0.05$ for frequency of session) than the other groups, and made significantly less uncorrected errors ($p<0.001$). Although this study was done on the learning of a physical skill and not the learning or retention of information, it does show that distributed practice is at least beneficial for procedural memory storage.

In 1987 both Bahrick & Phelps, and Dempster published experiments that specifically looked at the effects of distributed practice on the learning and retention of foreign language vocabulary (Bahrick & Phelps, 1987; Dempster, 1987). Bahrick & Phelps looked at the retention of Spanish vocabulary words over an 8 year period. In their study they recruited thirty-five participants who had to study 50 English-Spanish word pairs. In the initial training session, the participants were exposed to a presentation trail where each word pair was presented for a rate of 5 seconds per presentation. This was followed by a test trial in which the participants were prompted by the English equivalents to recall the Spanish target words. The participants had 10 seconds to pronounce the Spanish words. All the words incorrectly recalled were then presented again followed by a test trail. This alternation between the presentation and test trails continued until all the words were correctly recalled. All the following training sessions were the same as the initial session with the exception that they started with a test trail of all 50 word pairs instead of a presentation trial.

The inter-study intervals were 0 days, 1 day, and 30 days. After five to seven training sessions, the inter-study intervals (ISI) for most of the participants were changed. The original training lasted for on session, 8 months or 14 months, depending on the inter-study intervals of the participants. After a retention interval (RI) of *8 years* the participants did both recall and recognition tests. The results showed

that on average the 0 day ISI group recalled 6% of the words and recognized 71%, the 1 day ISI group recalled 8% and recognized 80% of the words, and the 30 day ISI group recalled 15% of the words and recognized 83%. An ANOVA on the probability of retention as a function of the inter-study interval and number of presentations during acquisition showed a statistically significant ($p<0.001$) main effect of the ISI (Bahrick & Phelps, 1987). Thus, the study of Bahrick & Phelps concluded that the optimum ISI for long term retention would be the longest possible interval before retrieval failures occur.

Incidentally, although the study focused on the ISI of the groups, it also showed that retrieval is essential. Thus, in effect, this study also supported the advantage of Spaced Recall Practice.

In the next study to be reviewed here, Dempster (1987) investigated the effects of variable encoding and spaced presentations on vocabulary learning. In a series of 5 experiments, Dempster compared the effects of context as well as distributed practice on vocabulary retention. 38 uncommon English words were used as the target vocabulary. The words were presented in a no-context condition, a one sentence context condition and a three sentence context condition. All five experiments had a no-context condition. Experiments 1 and 2 had a one sentence condition, and all experiments had a three sentence context condition. Each context sentence contained

the target vocabulary word as well as an explanation of the words meaning. In the no-context condition, the words only had explanations of their meanings. In experiment one, the words were only presented once. In experiment 2 to 5 the words were presented three times with 37 other words presented between each presentation. In experiment 3 to 5 a massed presentation condition was added (where the target words were presented directly after each other with no other target words presented between).

The results indicated no significant difference existed between the context conditions and the no-context condition ($p > 0.05$ for all the experiments) but there was a significant difference between the massed and spaced conditions for experiment 3, $F(1, 44)=6.57$, $p<0.025$, and experiment 4, $F(1, 44)=19.36$, $p<0.001$. The results of the series of experiments run by Dempster (1987) clearly indicated that there was a benefit to spacing the learning instances.

Dempster's experiment differed form that of Bahrick & Phelps (1987) in that Dempster's experiment focused on the micro level in that it took place in one learning session and the spacing was relatively small (the spacing of the learning instances took place within one learning session). Also, Dempster's experiment only had an immediate posttest (recall from working memory was controlled for by having the participants count backwards by threes for one minute). Thus the RI was close to 0. In

contrast, Bahrick & Phelps (1987) focused on the macro level in that their ISI's were in the order of days as well as having spacing within each learning session. Also the final RI was 8 years! Finally, while Bahrick and Phelps's experiment could also be seen as an example of Spaced Recall Practice, that of Dempster fell firmly within the confines of distributed practice.

Two factors that influence the effectiveness of distributed practice are the different kinds of relative spacing of the learning instances, called the *lag effect* (Cepeda et al,2008; Karpicke & Roediger, 2007; Karpicke & Bauernschmidt, 2011), and the actual length of the inter study intervals (Fritz et al 2007) (see figure 2.1).

Looking at the length of the inter study intervals, it has been shown that there is no silver bullet when it comes to the best spacing of learning instances. However, there are some broad guidelines. In general, as a coverall statement, the best spacing for long-term retention of information may be around 30 days (Bahrick, 1979). This is, however, an over simplification. The optimal spacing of learning instances depends on the length of the intended retention interval (Cepeda et al, 2008). For instance, if the intended retention interval is about 35 days, an optimum inter study interval might be 14 days. However, if the intended retention interval is 350 days, an optimum study interval might be closer to 25 days. There is a tradeoff between lengthening the inter study intervals though. Although it benefits long-term retention, it comes at a cost of

percentage of information successfully retrieved. Moving on to the lag effect, when it comes to how the learning instances should be spaced relative to each other it has been found that the relative spacing between the learning instances have less of an effect than the total spacing between all the learning instances (Karpicke & Bauernschmidt, 2011). For example, in Karpicke & Bauernschmidt's (2011) study of different relative spacings, they found that regardless of whether the spacing intervals were expanding, equal, or contracting, the test conditions with longer total intervals (the sum of the length of all the intervals), did up to 200% better than the test conditions with shorter intervals.

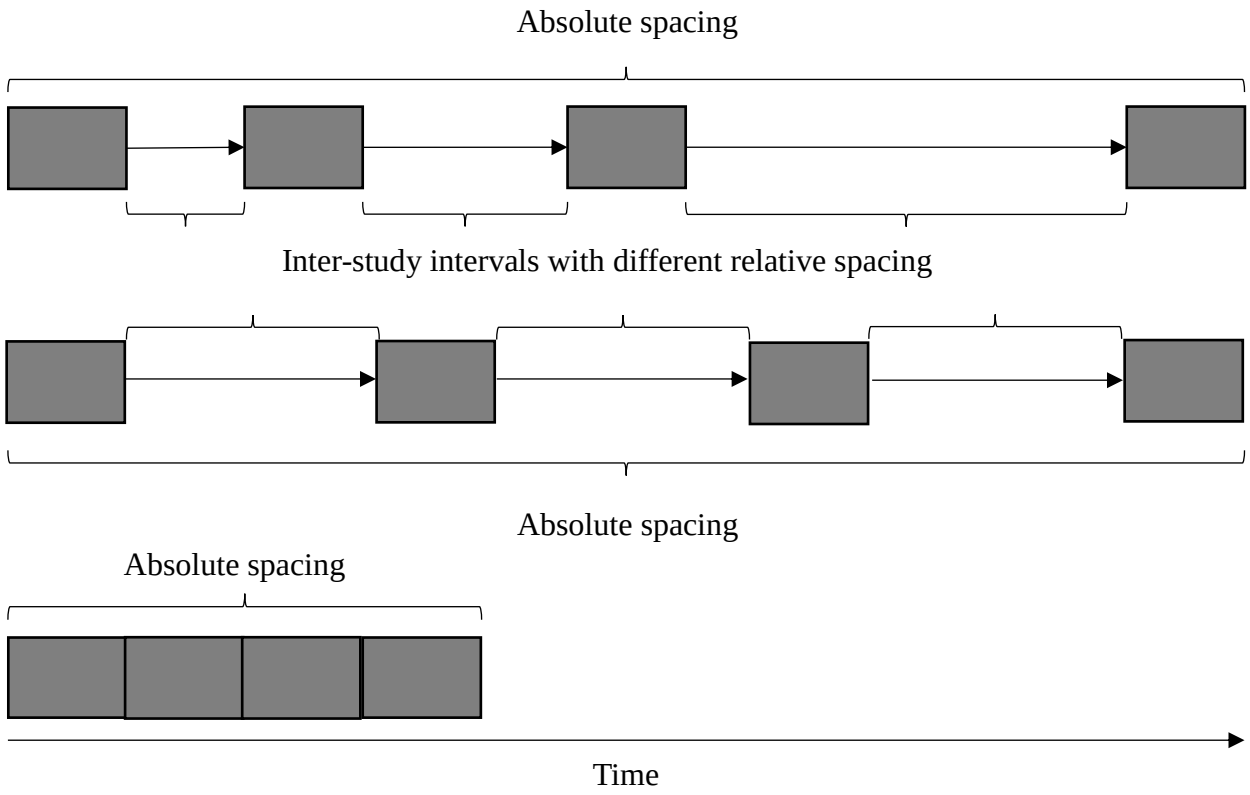


Figure 2.1 Absolute and relative spacing of learning instances

The above review is far from a comprehensive review of all the studies supporting the efficacy of distributed practice. For comprehensive reviews of distributed practice see Cepeda et al., (2006) and Dunlosky et al., (2013).

2.3.2 The Efficacy of Retrieval practice

When retrieving something from memory through recitation, tests, or simply recalling information, memory of that information is enhanced to a greater extent than simply restudying it (Gates, 1917; Spitzer 1939; Roediger, & Karpicke, 2006a). This is called the testing effect. The testing effect is one of the most well established psychological phenomena that is beneficial to remembering (Bjork & Bjork, 1992; Carrier & Pashler, 1992; Roediger & Karpicke 2006a; Roediger & Karpicke 2006b). The first empirical study of the testing effect was in done by R.S. Woodworth (1917). However, as Woodworth points out in his introduction with his reference to Sir Francis Bacon quote "…as if you read any work twenty times over, you will not learn it by heart so readily as if you were to read it but ten times, trying each time to repeat it, and when your memory fails you looking into the book," Bacon (1620, trans 1902), the knowledge that retrieving something from memory aids in memorization is much older than Woodworth's experiments.

Another early empirical study of the testing effect was done by Spitzer in 1939. In the test Spitzer used 3605 elementary school students to investigate the effect of

recall on memory retention, the effect of item difficulty on forgetting rates, and student ability and retention. Of importance to this literature review are the results of the investigation of retrieval on retention. In his experiment, Spitzer randomly divided the students per school into ten groups of about four hundred students each. Groups 9 and 10 were control groups. All the groups were given two articles to study and given 25 item tests on the first test. The first two groups were also given tests on the second article on the same day. The other groups were given their first test on the second reading either one, seven, 14, 21, 28, and 63 days after the initial study period. Group one was given a second test on the second reading one day after the first test, group two was given it seven days later, group three was given it 14 days later, group four was given it 21 days later, group five was given it 28 days later, and group six was given it 63 days later. The other groups didn't take a second test. Both group one and two also took a third test of the second reading on the 21st and the 63rd days respectively. The findings are briefly summarized in table2.2

Table2.2 *Results of the Spitzer (1939) studies in retention.*

groups	day						
	0	1	7	14	21	28	63
1	13.23	13.07	…	…	12.18		
2	13.20	…	11.83	…	…	…	10.74
3	…	9.56	…	8.93			
4	…	…	7.87	…	8.15		
5	…	…	…	6.97	…	7.10	
6	…	…	…	…	6.49		7.07
7	…	…	…	…	…	6.80	
8	…	…	…	…	…	…	6.38

Adapted from Spitzer, H.F. (1939) Studies in retention. *Journal of Educational Psychology*. 30: 641 –656

As can be seen from table 1, group one had a clear advantage over group three because of the additional recall they received on the day of the test. In general, looking at the results, it seems that the earlier the initial recall is spaced to the study of the test, the bigger the benefit to recall is. To control for repetition, group 10 was given a second test immediately after the first test on the day the students studied the

readings. No significant improvement was detected. However, because of the study design, it is not clear whether these results are because of the testing effect, the spacing effect, or mere repetition. If the groups could have been compared to a condition that had studied the reading at the same time as the other groups were doing the tests, the results could have been more robust.

To address this problem of whether the testing effect is just a manifestation of an additional study opportunity of a genuine effect in its own right, Carrie & Pashler (1992) devised a series of four experiments involving a total of 240 students at the University of California, San Diego. The first two experiments compared a pure study trial with a test/study trial to see which one was more effective. In both experiments, the subjects had to first study a set of paired associates. Then they were either shown both the stimulus and response for each pair together for ten seconds each (pure study trial, or ST), or they were shown only the stimulus for five seconds and then the stimulus and response for five seconds (Test Trial/Study Trial, or TTST). Finally they were given either an immediate a cued-recall test (5 minutes) or a delayed cued-recall test (24 hours). The main difference between experiment one and two were that in experiment one nonsense-word/number pairs were given as stimuli, and in experiment two English/Yupik (a Siberian Eskimo language) pairs were used. In both experiments the participants remembered more pairs from the TTST trial than from the ST trial.

Even though the students had five seconds more time to actually learn or review the words in the ST condition, the TTST condition still out performed the ST condition. If the testing effect were merely a result of another restudy opportunity then they ST condition should have outperformed the TTST condition (Carrie, & Pashler, 1992) or at least show no difference. Thus it could be concluded that the testing effect is indeed a genuine effect independent of just restudying of the same material. One shortcoming in this test is the lack of longer delayed post-tests. Compared to Spitzer's study, the delayed test is quite short. It would have been interesting to see if there were any differences in the rate of forgetting between the two methods and if the forgetting curves themselves were different.

In a later study to answer the question about whether the testing effect can be observed in an educational context, Roediger and Karpicke (2006) added evidence to the literature that the testing effect helps learning more than just restudying. In two experiments a total of 300 students from Washington University studied short prose about scientific topics. In experiment one, one-hundred-and-twenty of the students studied two readings. The experiment compared the restudy condition with the test condition using a within group design and compared the length of the delayed test (5 minutes, two days, or one week) using a between group study. The order of the learning condition and the readings were counterbalanced. The students were given

four seven minute periods to either study, restudy or take a test of one of the readings. Thus all the students studied one of the readings twice and studied the other one once and had a test on it. Between each session the students did multiplication problems for two minutes. Finally the students were given a recall test either 5 minutes, 2 days, or one week after the last study/test period. The results showed that on the 5 minute delayed test, the restudy condition did better than the testing condition (81% vs. 75%). However, on the two day and one week delayed tests, the testing condition did significantly better than the retest condition (68% vs. 75 % and 56% vs. 42% respectively). This study demonstrates that under certain conditions, the testing effect is beneficial to long-term retention. However, in this study, there was no feedback given to the testing condition. Thus, if something was incorrectly recalled, the information may be incorrectly encoded into longer term memory (Pashler, Cepeda, Wixted, & Rohrer)

The role of corrective feedback is very important for Retrieval Practice because it corrects erroneously recalled information. Two studies that look at the question of corrective feedback came out in 2005, and 2008 respectively. In 2005 Pashler, Cepeda, Wixted, & Rohrer conducted a study to ascertain if feedback helps with associated word-pair retention a week after the initial study session. An analysis of their results revealed that feedback had no effect on the answers that were correctly answered.

However, the feedback had an enormous positive effect on answers that were incorrect or left open in the initial test, and this positive effect persisted on the week delayed test (Palshler, et al. 2005). Later, Butler et al. (2008) investigated the effects of feedback not only of incorrect responses, but also low confidence correct responses on retention of general knowledge facts. They found that not only did the feedback group do significantly better than the no feedback group $t(29)=19.9$, $p<0.0001$, but low confidence correct answers were also retained significantly better if feedback were given than if feedback weren't given.

Looking at the testing effect as a whole, the literature shows that it is a beneficial effect in its own right, independent of just restudy, and in fact superior to it (Roediger, & Karpicke, 2006) However, in order for the effect to be fully realized and not to encode incorrect information into the long-term memory, corrective feedback should be given (Palshler, et al., 2005; Butler, Karpicke, & Roediger, (2008))

2.4.3 Efficacy of distributed practice and recall practice combined

As we have already seen above, distributed practice is the superior way of reviewing or repetition. However, the retrieval practice has also been demonstrated to be superior to mere repetition. Nevertheless, spaced repetition and feedback can both enhance retrieval practice (Roediger & Butler 2011). Thus, combining distributed practice with the retrieval practice should yield the best results (Roediger & Karpicke

2006b). As Nation (2001) points out, this can be done through the direct learning of words from flashcards. He asserts that learning from flashcards is important because it is efficient, it focuses on aspects of vocabulary knowledge that are not easily gained from other methods, and it allows for the spacing of repetitions. This last reason is important because with the help of flashcards, one can not only practice recall (testing) but these practices can be spaced, too.

One study specifically looking at the effect of spacing retrieval instances was that of Pyc & Rawson (2009). In their study, which consisted of two experiments, they used sort and long ISI to investigate the effect of retrieval effort on the retention of Swahili words. For the first experiment, they recruited 129 participants from Kent State University. The students learned 70 Swahili-English word pairs from 10 lists of 7 words each using a computer. All the word pairs were first presented in an initial study trial followed by retrieval trials. In the study trial, the participants were presented with the Swahili-English word pairs and given 10 seconds to memorize each pair. On all subsequent retrieval trials, the participants were given a cue word and they had to type in the target vocabulary. If an item was incorrectly recalled, 4 seconds were given to restudy the words.

The participants were divided into a sort ISI group and a long ISI group. The short ISI group studied the word pairs in blocks of 10. After the initial study trial, each

item was presented again for the practice trials until the item reached a criterion of either 1, 3, 5, 6, 7, 8, or 10 correct retrievals. Once the criterion was reached the item was dropped from the practice trials. When all the items in a block reached criterion, the next block was presented. This continued until all items in all 10 blocks reached the criterion level. The long ISI group learned the word pairs in blocks of 35. Also, in order to keep the length of the experiment reasonable the participants were given a maximum of 90 minutes to complete the blocks. Besides these differences, the procedure for the short and long ISI groups were the same. After completing all blocks the participants were given a 25 minute reading comprehension filler task as a distraction. Following that half the participants completed a cued-recall test that was the same as the practice trials but without feedback. This was the short RI group. The Long RI group did the same after an RI of one week. After performing a mixed factor ANOVA significant differences were found between the long and short ISI, $F(1,124) = 61.74$, $p < 0.001$, as well as between the long and short RI, $F(1,124) = 254.97$, $p < 0.001$.

The second experiment was the same as the first except for the following differences. First, the second experiment recruited 98 participants. Second, the computer recorded how long it took for the participants to start typing the answers once a cue was given in the retrieval trials. Finally, there was no one week delayed

posttest. Instead, short contracting and long contracting ISI groups were added to the experiment. With the contracting ISI groups the number of words in each block were gradually reduced so that the time between presentations were diminishing. This would theoretically make each retrieval easier than the one before. Although the fixed ISI groups seemed to outperform the contracting ISI groups, there were no statistical difference between the groups, $F(1, 92) = 3.31$, $p = 0.7$. However, there was still a significant difference between the long ISI groups and the short ISI groups, regardless of whether they were fixed or contracting, $F(1, 92) = 35.5$, $p < 0.001$.

Of interest to the current study is the fact that longer ISI had a significant difference on the retention of vocabulary for both the immediate posttest and the delayed posttest. This confirms that there is a benefit to combining distributed practice with retrieval practice. If the long ISI performed better than the short ISI, it can be inferred that the long ISI would also perform better than no ISI. In other words, Spaced Recall Practice would outperform massed retrieval practice.

2.4 Efficacy of Mnemonics and the Key word method

The Shorter Oxford English Dictionary describes Mnemonics as follows:

> The art of improving or developing the memory, especially by artificial aids; a system of precepts and rules intended to aid or improve memory.

In other words, Mnemonics is a group of techniques and principles used to aid in remembering. Mastropieri & Scruggs (2012) state, “A “mnemonic” then, is any procedure or operation designed to improve memory and/or remember something.” They continue to say that this term could refer to a very broad range of strategies from rehearsal strategies to transformational strategies. The strategies include verbal, imagery, and concrete forms. While some forms such as verbal forms could be quite simple, Mastropieri & Scruggs (2012) say that imagery forms or the concrete forms are more on the complex side of the scale. These forms involve transformation of the material intended to be learned into visual imagery in the learners mind or into concrete pictures etc. The earliest evidence of such a transformational strategy comes from the *ad Herennium* (*Rhetorica ad Herennium*) around 86-82 B.C. and involves the *loci* method (Yates, 1966) also known as the memory palace technique. Transformational mnemonic techniques include letter strategies and the pegword method among others.

One mnemonic technique that has received quite a lot of attention is the Keyword method. Refined by Atkinson and Raugh (Atkinson, 1975), the key word method involves a two-step process. First, the L2 word is associated with a phonetically similar key word that the learner is already familiar with (either an L1 word or a known L2 word). In this way the sound of the word is linked to an already established

schemata in the brain. Atkinson calls this the *acoustic link*. Next, a mental image is created where the key word interacts with the target L2 word so that the meaning of the target word is also linked visually to an existing concept. Atkinson calls this the *imagery link*. An example adapted from Chen (2006) is that of the noun "miss" and the Mandarin keyword 迷死人. Miss sounds like the first two characters of 迷死人 (mísǐrén), which means very charming in Taiwanese Mandarin. Those two words make an acoustic link, linking the word to be learned with an existing L1 word. Imagining a very charming teacher or lady with a nametag "Miss so and so" could create the imagery link, thus linking the L1 one word with the L2 word meaning.

2.4.1 Atkinson and Raugh's original experiments

In a series of experiments on English native speakers learning Spanish and Russian, Atkinson and his associates found that the keyword method that they employed was a more effective method for memorizing new vocabulary in an L2 (Raugh, Schupbach, & Atkinson,1977; Raugh & Atkinson 1975; Atkinson & Raugh 1975; Atkinson,1975).

The first series of experiments

The first series of for experiments was on English learners of Spanish (Raugh & Atkinson, 1975). In all four studies, the participants were native English speaking Stanford University students with no prior knowledge of Spanish. None of the

participants participated in more than one of the studies. The keywords that were used were selected according to three criteria; they had to sound as much as possible like the target word or part of the target word, it had to be easy to form an imagery link between the keyword and the target word, and the keywords had to be unique (there couldn't be a key word associated to more than one target word). The first experiment involved the pre-learning of the keyword target word pairs before linking them with a mental image to the L1 words. This was done to see if mental imagery could be used to link the word pairs after the first step. Forty participants were used, and after the first step they were divided into an experimental group and a control group. The experimental group was instructed to use mental imagery to link the key words with the target words whereas the control group was told to directly link the target words with their L1 equivalents using rote rehearsal. Participants studied words presented in both written and oral form. After the study session the participants were given three immediate post-tests. They tested oral Spanish to written English, written Spanish to written English, and spoken Spanish to written English. The results showed for the first test the experimental group outperformed the control group with average scores being 88% to 28% respectively ($t=14.74$, $p<0.001$). For the second test the results were similar with the groups scoring 88% and 32% respectively ($t=11.56$, $p<0.001$). For the Keyword Spanish comparison, there was no significant difference in the

results.

The second experiment was basically the same as the first experiment with the exception that the Spanish words were only presented orally and after the keyword target word pairs were practiced the L1 words were linked with the keywords while listening to the Spanish pronunciation. Thus the keywords were not linked to the L1 words in isolation from the target words as in the first experiment. Again, two immediate posttests were given. The first tested the oral Spanish to written English associations, and the second tested the oral Spanish keyword associations. There were significant differences between the results of the experimental group and the control group on both tests. The results for the first test were 50%, and 30% respectively for the experimental and control groups ($t= 3.2$, $p < .01$) and 60% and 53% respectively for the two groups ($t= 2.9$, $p< .02$) on the second test.

The third experiment differed considerably from the first two. First, it was run over three consecutive days and a vocabulary of 120 words was used. Second, the imageability of the keywords was controlled. Third, a posttest was given two days after the completion of the learning sessions and a delayed posttest was given about one month later. Finally, this experiment was a within group design with two conditions; in the first condition, the participants memorized the words with the aid of given keywords and in the second condition, the participants memorized the words

anyway they wanted but without the aid of a given key word. The results of the first posttest were 54% for the experimental condition and 45% for the control condition (paired $t=4.1$, $p<0.001$) and for the delayed posttest it was 43% and 35% respectively (paired $t=3.5$, $p< 0.01$). Furthermore they found that words with a high imageability were more likely to be remembered than words with a low imagebility in the Keyword condition (0.56 probability compared to 0.50 probability), whereas in the control condition there didn't seem to be a significant difference (0.44 probability and 0.45 probability, respectively).

The final experiment was the same as experiment 3 except for two things; firstly, a free choice condition was added where the participants could choose to use a given keyword or not, and secondly, there was only one posttest, one day after the completion of the treatment. The results for the first posttest or the free-choice, keyword, and control conditions were 59%, 57% and 50% respectively, $F(2,48)= 6.94$, $p< 0,005$. Although significant differences emerged between both the experimental conditions and the control condition at the $p=0.05$ level, the difference between the experimental conditions wasn't significant.

The results for both the first and second experiments demonstrated that the keyword method, if applied in a similar way as Raugh and Atkinson did, should yield better results than rote repetition. Furthermore, the third and final experiment also

yielded positive results for the Keyword method compared to a free choice of methods employed by the participants. However, most of the results (with the exception of the delayed test for the third experiment) only relied on immediate posttests or delayed posttests that were done one or two days later. Thus, the long-term benefits of the keyword method could not be conclusively demonstrated with these experiments. Furthermore, although the recognition abilities were tested in all the experiments, their recall abilities weren't. In other words, the Spanish words were given and then the English translations (the participant's native language) had to be recalled. Thus the efficacy of the keyword method to promote productive use was untested. Finally, some of the subjects had studied other Romance languages. This might also have had an influence on the results, making it easier for the participants to use of find cognates.

The second series of experiments

Some of these problems were addressed in their next experiment (Atkinson & Raugh 1975). In this experiment 52 students from Stanford University participated in learning 120 Russian words on three consecutive days. They were all native English speakers and none of them had studied Russian before or participated in the previous studies. Half of the students were male and half were female. This was important, because in the first series of experiments Raugh & Atkinson (1975) had found that the

female students were significantly better at learning the vocabulary words. All the keywords paired with the Russian target vocabulary were ranked according to their imageability and their abstractness. The students were randomly assigned to a control group and an experimental group with the condition that both groups contain the same amount of males and females. The experimental group were trained to use the keyword method and instructed to use it to learn the words during the treatment phase of the experiment. The control group was instructed to use any method they liked to learn the target Russian words and their English equivalents. The participants wrote an immediate posttest the day following the last treatment and a delayed posttest between 30 and 60 days after the immediate posttest.

Two significant results of the first posttest were that the keyword group outperformed the control group, $F(1,48) = 35.8$, $p.< 0.001$, and females outperformed males, $F(1,48) = 5.9$, $p.< .025$. On the delayed test, the keyword group again outperformed the control group. The probability of a correct response for the keyword group was 0.43 and for the control group it was 0.28., respectively. The question of whether the keyword method's effects were confounded by language cognates in the first series of experiments seem to have been answered by the Russian-English experiment. As there are not many cognates shared between the languages and the results were positive in favor of the keyword method, it could be safely assumed that

if there is a confounding effect because of cognates in the first series of experiments, it is not significant to the results or conclusions. This experiment also laid to rest the question of the long term effectiveness of the keyword method as used by Atkinson & Raugh in their experiments.

When it came to the imageability of the vocabulary, there was no significant difference between the high and low imageability words in the keyword group. However, there was a significant difference in the control group, $F(3, 25) = 3.1$, $p. < .05$. This is striking, because in experiment 3 of Raugh & Atkinson (1975), the imageability of the keywords seemed to have an effect on the keyword condition and not the control condition. At first these results may seem contradictory. However, in Raugh & Atkinson (1975) experiment 3 the keywords' imageability were tested, whereas in Atkinson & Raugh (1975) the English translations of the target words were tested. Therefor it would make sense that in the Spanish-English experiment the keywords have more of an effect because of two reasons. First, the keywords are used to form a mental image connection with the target word. Therefore, the easier it is to imagine the keywords, the easier it should be to create such an image in one's mind, thus facilitating the learning of the word. Second, since the control group didn't use keywords to learn the target words, there is no way that those words could, even as a confounding factor, affect the results for the control group.

More interesting though are the results of the Russian-English experiment. The results for the control group could be explained by the fact that more concrete words seem to be easier to learn than more abstract words. And since they didn't have keywords to facilitate the learning, the ease of learning of the target and English word pairs would have a greater influence on them. However, with the keyword group, the keyword acts as a bridge between the L1 and target word pairs. If the keyword were easier to learn because of its higher concreteness, then it could be that it would have a bigger effect on the ease of learning the target vocabulary than the L1 and target vocabulary have.

In conclusion, the experiments ran by Raugh and Atkinson showed that the keyword method does, under certain conditions, improve both the immediate and long-term retention of vocabulary recognition abilities. It also showed that there is a significant difference between the effectiveness of the keyword method between male and female participants. This implies that for reliable data on the keyword method, the sex of participants in any experiment regarding the keyword method needs to be controlled for. The experiments also showed that the concreteness of the keywords used in the experiments is a significant factor to its success. Finally, Raugh & Atkinson's pilot studies hinted that, although it is better for the keywords to be given to the participants than for them to come up with their own keywords, it is better for

the participants to create their own mental image or verbal connection than it is for these to be provided to them. However, the experiments do not conclusively answer the question of the efficacy of the method to aid in the productive acquisition of the target vocabulary.

2.4.2 Empirical studies related to the Keyword method

Although Raugh & Atkinson concluded that it may be better for students to generate their own imagery links, Pressley, M. & Levin, J. R.'s (1978) study of second and sixth-graders learning Spanish vocabulary words suggests that this assertion may not hold true for younger learners. In their study they compared three imagery conditions with a control condition. In the first condition the students were given pictures of the keywords and target words interacting. In the second condition, the pictures of the keywords and target words were separate and the students had to come up with their own interactions. In the final imagery condition, the students were only given the keywords and the target words. The second graders in the first condition remembered the most vocabulary, followed by those in the second condition. Those in the third condition were not significantly different from the control condition. The sixth graders performed the same or better than the control group in all conditions. Although the results for the sixth graders are mixed, it is clear that the younger learners (the 2nd graders) benefitted from having imagery links provided to them. This

may be because at younger ages learners my find it too mentally demanding to create these imagery links themselves.

When it comes to presentation rates Hall, Owens & Willson (1987) showed that longer presentation rates are essential for the keyword method, but may be counterproductive to other methods. They ran two experiments comparing the keyword method with paired associate learning (PAL) at different learning intervals. In the first, words were either presented once at an 8 second interval or twice at a 4 second interval. It was found that the PAL group did better at the 4-second interval than at the 8-second interval. However, the keyword group was unable to complete the experiment because the presentation rate was too fast for them to form imagery links.

In the second experiment, 20 participants who had to study rare English words were assigned to three conditions, namely, 3X 3-second PAL presentation, 1X 9-second PAL presentation, and a 1X 9-second keyword method presentation. They were presented with 22 rare English words. The results showed that the 3X 3 second PAL group scored the best on the immediate posttest, followed by the Keyword group, and the 1X 9 second PAL group did the worst. This study showed that if the presentation rate is controlled for different learning methods, it may produce a distorted result. For example, the PAL group did better at 3-second presentation rates

than at 9-second presentation rates. Also, the keyword group needed more time to implement the technique.

However, the number of repetitions given to the 3 second PAL group may have confounded this study. In order to clear up the impact the repetitions had on the results, both the 9-second PAL and the keyword conditions needed to have the same amount of repetitions as the 3-second group. In fact, Pressely (1988) reviewed four studies challenging Hall, Owen, and Wilson's assertion that the timing of presentation biased results in favor to the keyword method. The most convincing study reviewed was Pressley's own study where he modified Hall's methodology to get rid of the confounding factors by having one keyword method with a 1x9 second presentation rate and one with a 3x3 presentation rate (Pressley, 1987). When properly instructed and trained to use the keyword method, Pressley found two effects, the 3x3 presentation rate was superior to the 1x9 presentation rate and the keyword method was superior to the control condition. Thus, it seems that, if properly instructed, the keyword method could in fact be successfully combined with a distributed practice method to yield superior outcomes than either just distributed practice or just the keyword method.

2.4.3 Combining the keyword mnemonic with Spaced Recall Practice

A more recent study directly comparing the Keyword method with a form of

distributed practice (expanding schedule retrieval practice) is that of Fritz, Morris, Acton, Voelkel, and Etkind (2007). They ran a series of three experiments in which they compared the retrieval practice method, the keyword mnemonic, and a combined retrieval keyword group with a control group. The first two experiments only compared the keyword mnemonic and the retrieval practice method with control groups and tested their recognition of the foreign vocabulary.

In the first experiment, both the experimental conditions outperformed the control significantly ($p < 0.01$). However, there was no significant difference between the two experimental groups after either a 3-minute delay or a three-day delay. Fritz, et al (2007) postulated that a ceiling effect might have been in play in the first experiment.

Thus experiment two was devised to control for this ceiling effect by increasing the number of words studied, shortening the presentation times and increasing the delay of the first posttest to 24 hours. Furthermore, they lifted any restrictions on the way the control group could study the new words. The results again showed a significant difference between the two experimental groups and the control group ($p < 0.01$), with the keyword group doing slightly better than the retrieval practice group. Again, there was no significant difference between the two experimental groups.

In the third experiment, a combined keyword retrieval practice method was

added to the experiment, the control group was instructed to use elaborative learning, both recognition and production of words were tested, and a one week delayed test was added. In all four posttests (immediate recognition, delayed recognition, immediate production, and delayed production) the combined method, and the two individual test methods were significantly better than the control condition (p, 0.001). In both the immediate and delayed recognition tests, there were no significant differences between the experimental conditions. However, in the production tests, the retrieval practice condition was significantly better than the keyword method condition in both the immediate (p=0.012) and the delayed (p=0.028) tests. Furthermore, there was so significant advantage to the combined method condition over the retrieval practice condition. Also, the combined condition actually fared slightly worse than the retrieval practice condition, although the difference was non-significant.

This study certainly raises a few eyebrows because, if the results could be reproduced in a longer-term study, it would suggest that when it comes to initial learning of novel words, the use of a complex mnemonics such as the keyword method is unnecessary. Also, although it would be expected that a combined condition would result in added benefits from both methods because they rely on different cognitive mechanisms (Fritz, et al, 2007), the results from this study points to the

contrary.

An interesting observation made by Fritz et al is that the difference between the test scores on the immediate recognition tests and the immediate productive test were not significant for the combined condition. This is different to what one might expect, and also different from all the other conditions in the experiment. However, this advantage seems to be lost by the one week delayed test.

One point that should be raised though is that the final experiment used a within group design and the various conditions were not counterbalanced for order of use. This, as Fritz et al (2007) admits, could be a confounding factor influencing the combined practice condition. Thus, a similar experiment that compares the various conditions using a between group design might control for this confounding factor. Also, as suggested by Fritz et al, it would be interesting to see if the benefit to the productive performance produced by the combined practice could be retained for longer periods if retrieval practice combined with the keyword method is extended over a period of a month or even longer.

CHAPTER THREE

METHODOLIGY

As already stated, this research aimed to examine the effects of combining the Keyword mnemonic with Spaced Recall Practice on vocabulary learning. A within group experimental design was used to control for the fact that participants came from different classes. Forty-nine 8th graders started the experiment. However, because of the extended nature of the experiment and attrition, only forty-two participants completed the experiment.

The treatment consisted of three stages. In the first stage, the participants were taught how to use the keyword mnemonic. In the second stage, the participants were first sorted into participant groups. Then each participant group studied the target vocabulary once a week for four consecutive weeks. In the third and final stage, a third of each participant group did the immediate posttest, a third from each group did the one week delayed posttest, and the remaining participants did the four week delayed posttest (See appendix D)

Twenty-one target English-Chinese words pairs were divided into three lists of 7 words each. Each list was controlled for difficulty so that all three lists were of equivalent difficulty. Each participant group only studied two of the lists using the

combined method for one list and Spaced Recall Practice for the other. The combined method consisted of the keyword mnemonic combined with Spaced Recall Practice. The methods used to memorize the sets of words were rotated for each consecutive group so that, in the end, each method had been applied to each set of vocabulary one time by one of the groups. In addition, each participant group was not exposed to the remaining set of seven words in the rotation. These words were included in the posttests, though, and acted as the control.

The following sections will describe the participants of the study, the vocabulary words, the vocabulary learning booklets, the instruments used to measure the dependent variables, the dependent and independent variables, the treatment procedure, the data collection procedure, and finally the data analysis.

3.1 Participants

In order for the participants to be representative of the population of EFL junior high students in Taiwan, a few selection criteria were implemented. First, any students that grew up in an English bilingual environment were excluded from the study. This is because they may already have had a higher number of exposures to the target vocabulary used in this experiment than what is normal in the population the test participants should represent. Second, for the same reason as just mentioned, any students that had stayed for an extended period in a foreign country were also

excluded from the experiment.

Forty-two 8th grade junior high students from a private school in the Taichung area completed this study. The students came from six eighth grade homerooms. All the students had five hours of ESL class per week as well as five hours of regular English class for a total of ten hours of English a week. The regular English classes were those stipulated in the curriculum guidelines by the M.O.E. The ESL classes differed from the regular classes in so far as native English speakers taught them and the classes were English only environments (with necessary exceptions). Furthermore, a strong communicative approach was taken, so, in class, there were a lot of pair work and group work activities where information needed to be exchanged and tasks needed to be fulfilled. Also, the ESL classes were smaller than the regular English classes. Consequently, there were 12 ESL classes whereas there were only 6 regular English classes; the students were recombined from their six homeroom classes into the twelve ESL classes based on the students' English ability. The ESL abilities of the students were determined by a combination of their midterm tests and their end of semester exams. 15 of the participants were female and 27 were male. Of the forty-two students recruited for the study, twenty-four were intermediate level students and eighteen were high-intermediate level students.

3.2 Target Vocabulary Words

Twenty-one words were selected from the High-intermediate GEPT test vocabulary list (retrieved from: www.lttc.ntu.edu.tw/wordlist.htm). The high-intermediate GEPT test vocabulary list was chosen because it was assumed that the majority of junior high students would not be familiar with these words. In support of this assumption, between 2012 and 2014, only 6% of the high intermediate GEPT test takers were junior high students. Furthermore, only 3% of the high intermediate test takers between 2013 and 2014 were in the 12-14 year old age range (retrieved from: www.lttc.ntu.edu.tw/results.htm). Thus, it can be inferred that only a small percentage of the high intermediate test takers are grade eight students and although this does not directly show that the students won't know any of the words, it does indicate that the general level of English for junior high students is lower than that of the high intermediate GEPT.

In order to control for word type, difficulty, and concreteness of the words, a few selection criteria were introduced. Firstly, only nouns were used in this experiment. This is because, as Nation (2001) states, word type may have an influence on the learnability of the words. Secondly, to control for the difficulty of the words because of the number of syllables, only one to three syllable words were used; twelve were one syllable words, six were two syllable words, and three were three syllable

words. Finally, the concreteness ratings of the words were controlled for (Brysbaert, et al. 2014). The concreteness of a word refers to how tangible a word is and thus the likelihood that this word has a corresponding image (Clark & Paivio, 1996). The rating system used in this experiment, produced by Brysbaert, et al. (2014), ranges from 5 to 1. The highest rating of 5 is given to concrete words for which an image can easily be imagined such as apple, flower, or frog. On the other side of the spectrum, a rating of 1 is given to much more abstract words which are more difficult to associate with an image such as essentialness and spirituality.

The selection procedure was started by sorting the High GEPT vocabulary list into word classes using excel and then only retaining the nouns. After that, all words with four or more syllables were removed. Next, all the words with concreteness ratings of less than 4.5 were removed and the remaining words were saved to the experimental wordlist pool. Following this, the experimental wordlist pool was randomized into groups of one, two and three syllable words. Then the top 21 words were selected in a ratio of 4:2:1 from the one, two and three syllable groups respectively. These words were then presented to a group of experienced Taiwanese ESL teachers who came up with Mandarin Keywords for the words. Any words for which it were impossible to find an appropriate keyword were dropped from the 21 word list and new words from the corresponding syllable word lists were randomly

selected to replace them. Only 21 words were chosen due to time constraints. Only 20 minutes were available each week. It was calculated that if 8 seconds were spent on each page, it would take just over 11 minutes to complete each booklet. This combined with getting the class ready for the treatment limited the amount of words used.

Later, after results from the pilot test, four final words were dropped from the 21 word list because they didn't discriminate enough. Four new Keywords were created for two one syllable words and two three syllable words. These Keywords were created by the grade seven class that was also the class of the participants of the pilot study. The students were given a list of 10 two syllable words and six three syllable words from the experimental wordlist pool. They were given 15 minutes to come up with keywords for the vocabulary. This was done as part of an exercise to review and reinforce the keyword mnemonic. Following this, the words were presented to the Taiwanese ESL teachers and they chose the four they found the most appealing (see appendix A for a list of all the vocabulary word and their keywords).

3.3 Vocabulary Learning Booklets

The vocabulary learning procedure will be discussed in section 3.6.3. In this section only the booklets themselves will be discussed. Once a week for the four vocabulary learning weeks, each participant received 84-page booklets containing the

14 target vocabulary words that their group was studying. The cover stated the group number, the participant number, and the week number. In the first week, the first 28 pages of the Vocabulary Learning Week One booklet were collectively known as the learning phase. These pages contained Learning page (LP) and Retrieval page (RP) pairs. In the learning phase, the LP's always preceded the RP's (Figure 3.1 & Figure 3.2). If the page pair was from the Spaced Recall condition, the LP contained one English word and a Chinese equivalent, and the RP contained only the Chinese equivalent with a space to write the English word (Figure 3.1). If the page pair belonged to the Combined Condition, the LP contained one English word, a keyword, a sentence suggesting how to imagine the keyword and the meaning interacting, and a Chinese equivalent (Figure 3.2). Again, the RP contained only the Chinese equivalent with a space to write the English word (Figure 3.2). The next 56 pages were collectively known as the recall phases and consisted of two recall phases. Each recall phase was the same as the learning phase except that the RP's preceded the LP's (Figure 3.3 & Figure 3.4). All 14 target words were presented in a random order in each phase. The following three weeks' booklets were the same as the first booklet except that they contained three recall phases instead of a learning phase and two recall phases.

Figure 3.1. Example of a Spaced Recall Condition Learning Phase word pair where the Learning Page precedes the Recall Page.

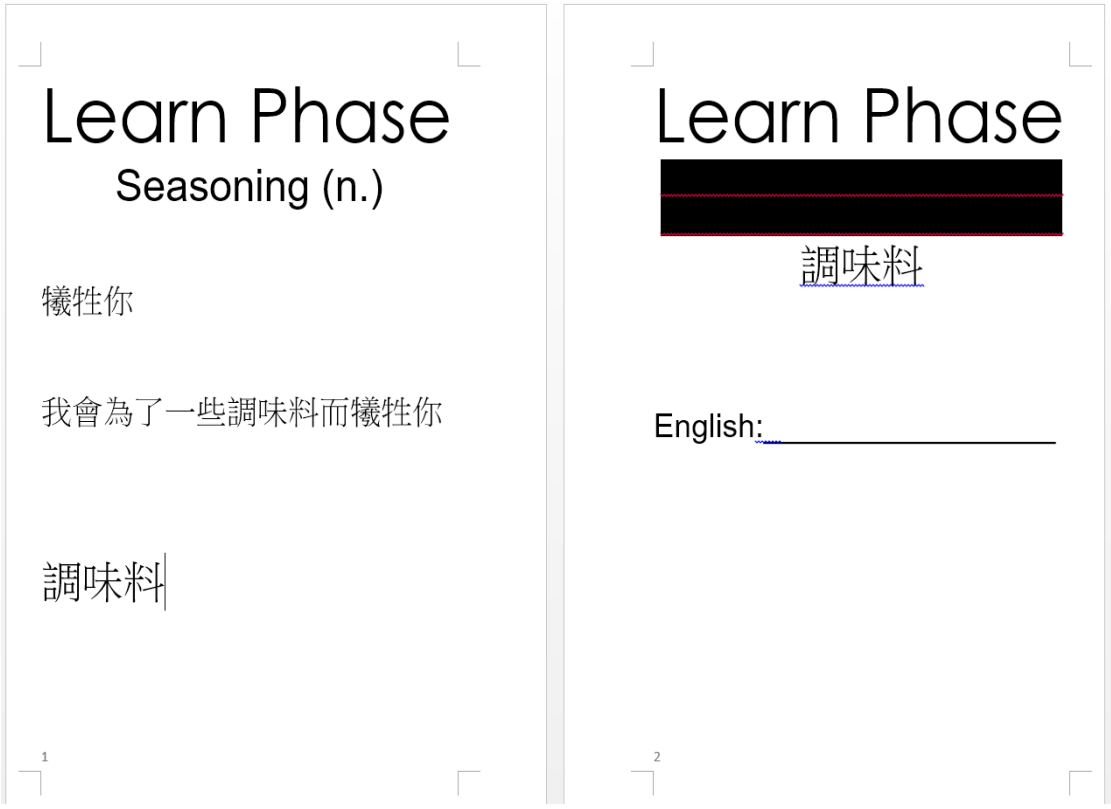


Figure 3.2 An example of a Combined Condition Learning phase page pair where the Learning Page precedes the Recall page

Figure 3.3 An example of a Spaced Recall Condition Recall Phase page pair where the Recall page precedes the Learning Page.

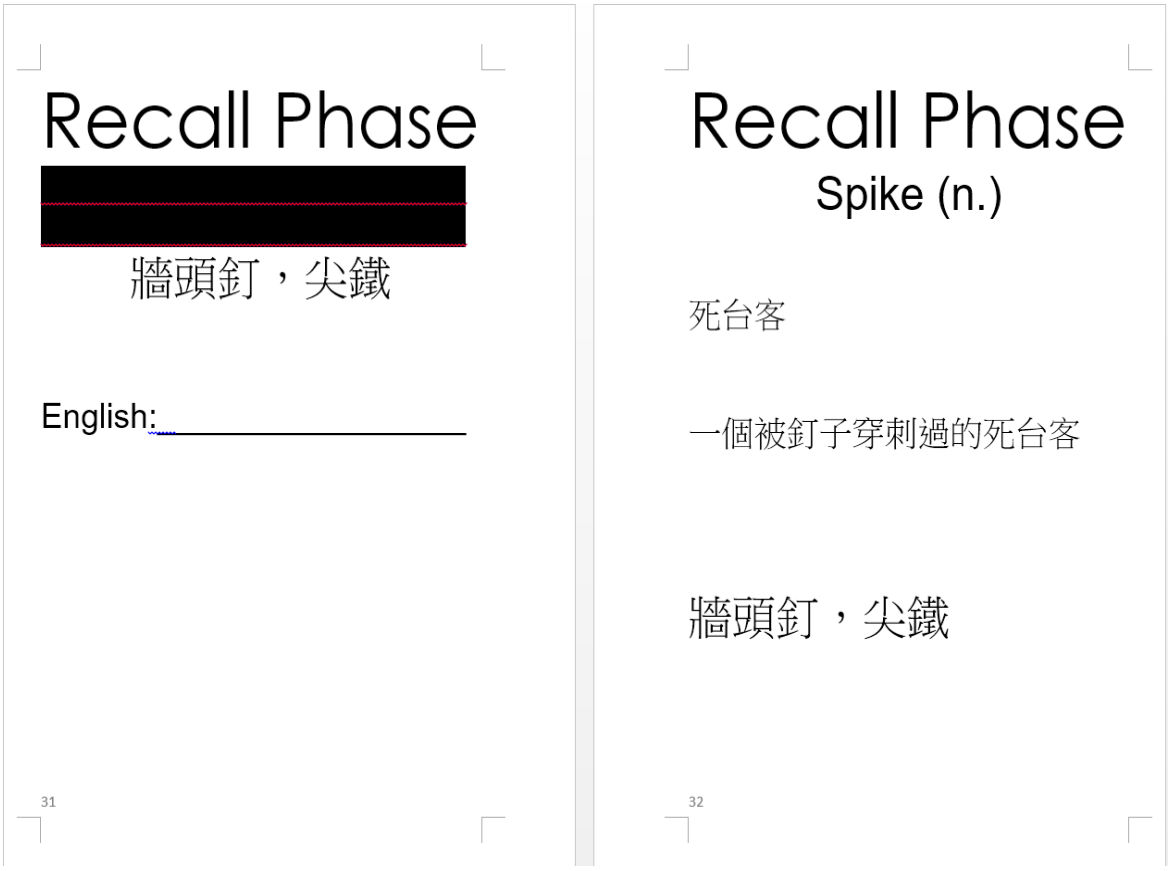


Figure 3.4 An example of a Combined Condition Recall Phase page pair where the Recall Page precedes the Learning Page.

3.4 Instruments

The instrument used to measure the vocabulary levels of the participants in this study was the Mandarin version of the Vocabulary Size Test of Paul Nation (Nation & Beglar, 2007) (see Appendix B). The test was converted from the Simplified Mandarin version to Traditional Mandarin characters using online conversion software (http://www.chinese-tools.com/tools/converter-simptrad.html). The Vocabulary Levels Test (Nation & Beglar, 2007) was used for the purpose of dividing the participants into the three participant groups.

The instrument used to measure the vocabulary retention (The Mandarin to English Productive Vocabulary test) contained the Mandarin equivalent words as the prompts of the English target vocabulary as well as a space to enter the English target vocabulary. The Mandarin to English Productive Vocabulary test was administered at three time points (immediate posttest, one-week delayed posttest, and four-week delayed posttest). This was used to test the productive vocabulary of the participants. All the 21 prompts as well as the answer spaces were printed on one page so that they could all easily be scanned with a glance (see Appendix C).

3.5 Variables

The main *construct* under investigation in this study was the *Productive English vocabulary* of the junior high students participating in this study. The *dependent*

variables representing these constructs were the amounts of English vocabulary retained by the participants after *RI*s of no time, one week, and four weeks. The *operational definition* of the *dependent variable* in this study, the *productive English vocabulary,* is defined as the amount of *L2* vocabulary correctly translated from the given *L1* equivalents in the posttests.

The *independent variables* are the treatment conditions and the retention intervals. The treatment conditions are the Spaced Recall condition and the Combined Method condition (the combination of the Keyword Mnemonic with Spaced Recall Practice), as well as a control group. The treatments differ in that the Combined Method condition incorporates the use of keywords in a spaced recall schedule, while the Spaced Recall Practice condition solely relies on the rote memorization of associated word pairs repeatedly spaced over time. In the control condition, participants won't study the words that will be tested. The three retention intervals are the no RI, the one week RI, and the four weeks RI. The only difference between the retention intervals is the amount of time between the last intervention session and the posttest.

3.6 Treatment procedure

The instructional procedure started with teaching the participants how to use the Keyword Mnemonic. After this the participants wrote the Vocabulary Levels Test

(Nation & Beglar, 2007). They were then sorted into their participant groups based on the results of this test. Following that, the participants memorized the target vocabulary words using the Combined Method and Spaced Recall Practice. This was done by working through the vocabulary learning booklets at a controlled pace.

3.6.1 Keyword Mnemonics method

As stated above, the treatment procedure started with instruction in using the keyword mnemonic. This instruction method was adapted from the original experiments run by Atkinson and Raugh (1975). The Keyword Mnemonic learning sessions took place twice within one week. All the participants were introduced to the Keyword Mnemonic, and instructions were presented on how to use the keyword method via an overhead projector. Also, the participants used their ESL workbooks for exercises they had to do related to the Keyword Mnemonic.

The training was divided into the following six steps:

Step one: the participants were taught to make an *acoustic link* where they had to associate the target English word with a similar sounding Mandarin word (the Key word). They were shown an English word with its acoustic link on the projector and asked to write it down in their work books. They were then told to take a moment to recall the link in their minds with their eyes closed. For example, they were given the word *collection* and the keyword 可樂神.

Step two: the participants were taught how to make the *imagery* link, where they have to associate the keyword to its meaning via an image where the two concepts interact. The participants were shown an explanation of how the keyword and the meaning could interact in an imaginary interactive image. Finally, the Mandarin equivalent of the English target word was also shown. The students were instructed to take a moment to look at the explanation and then to try and visualize the image in their minds without looking at it. For example, for the keyword, 可樂神, they were given a moment to close their eyes and imagine the "cola god' looking at his collection of cola bottles. They were then told to recall the Mandarin meaning from the imagined interactive image and then the target English vocabulary.

Step three: On the next page of their notebook they were only given the explanation of the interactive image. They were then instructed to first recall and write down the L1 meaning, then the keyword, and finally the L2 target word. Feedback was then immediately given.

Step four: the above procedure was then repeated twice more in the same way with other examples.

Step five: the students were given a practice English word with its Chinese equivalent, the associated keyword and a suggested explanation of how to imagine the interactive image. They were then told to use the procedure they practiced in the first

four steps to memorize the English-Chinese word pair. They were then given the Chinese prompt, and told to visualize the interactive picture, and recall and write down both the keyword, and then the English target word. Finally, they were given immediate feedback. This was then repeated once more in exactly the same way, and then two more times in the same way except that they didn't have to write down the keyword.

Step six: after a delay of between 1 and 3 days, step five was repeated. However, this time the students were limited to 10 seconds to make the acoustic link and the imagery link. They also only had to recall the target English vocabulary words. This addition to the original method was done in order to gain the advantages of the spacing effect, thus ensuring that the students understood how to use the Keyword Mnemonic.

3.6.2 Sorting of participant groups

All the participants took a modified version of Nation's Vocabulary Levels Test before the experiment began (see Appendix B). As stated above, the results of this test were used to help sort the participants into groups with a similar spread of language ability.

When it came to confounding factors, one that needed to be controlled was the gender of the students. The reason for this is when it comes to the keyword method,

gender has been shown to be a confounding factor (Atkinson & Rough, 1975). In their series of experiments, Atkinson and Rough (1975) found that females significantly outperformed males, $F, (1, 48) = 5.9$, $p < 0.025$. Furthermore, in a study of children aged between 9 and 15, differences in the neural processing of language have been found between males and females (Burman, Bitan, & Booth, 2008). To control for this, each group required similar numbers of girls and also similar numbers of boys. In other words, if one group had seven girls, both the other two groups needed six or seven girls. The same held true for the boys.

Another potential confounding factor was vocabulary size. As already mentioned, vocabulary size is a good predictor of language ability (Alderson, 2005; Laufer, 1997; Laufer & Goldstein, 2004; Nation, 2006; Schmitt, Jiang, & Grabe, 2011). Furthermore, language ability, and in particular, reading ability, has been shown to influence the rate of vocabulary acquisition (Swanborn & de Glopper, 2002, Laufer & Paribakht 1998). Thus it is reasonable to assume that students with a larger vocabulary size may also acquire new vocabulary at a different rate than those with a smaller vocabulary size. Consequently, to control for the existing vocabulary knowledge of the students, each intervention group needed a similar spread of student vocabulary sizes.

To control for the above-mentioned confounding factors, the vocabulary sizes of

the students first had to be determined. The students were tested on their vocabulary levels using the Vocabulary Levels Test (Nation & Beglar, 2007). Table 3.1 shows the participants' performance on the Vocabulary Levels Test. Of all the participants who finished the study, one reached the 1200 level, one the 1300 level, and one the 1500 vocabulary level. Four participants reached the level of 1700 words, five the 1800 level, five the 1900 level, and five the 2000. Only two participants reached the level of 2100 words, four the 2200 level, three the 2300 level, three the 2400 level, one the 2500 level, one 2600 level, two the 2700 level, and two the 3000 level. The best two participants reached the vocabulary levels of 3200 and 3500 respectively. The median Vocabulary Level achieved by the participants fell in the 2000 words level. In other words an equal number of participants have scores lower and higher than this number.

Figure 3.5 shows the distribution of participants reaching each of the vocabulary levels on the test. The range is from the 1200 vocabulary level at the low end to the 3500 vocabulary level at the high end. The graph has three modes of 5 participants each at the 1700 vocabulary level, the 1800 level and the 1900 level. Interestingly, fifty percent of the participants fall within the 1700 to 2100 vocabulary levels. The graph follows a rough positively skewed distribution with a dip of 2 participants at the 2100 level and another dip of one participant each at the 2600 and 2700 vocabulary level.

Table 3.1 *Vocabulary levels of the participants*.

number of students	vocabulary level
1	1200
1	1300
1	1500
4	1700
5	1800
5	1900
5	2000
2	2100
4	2200
3	2300
3	2400
1	2500
1	2600
2	2700
2	3000
1	3200
1	3500

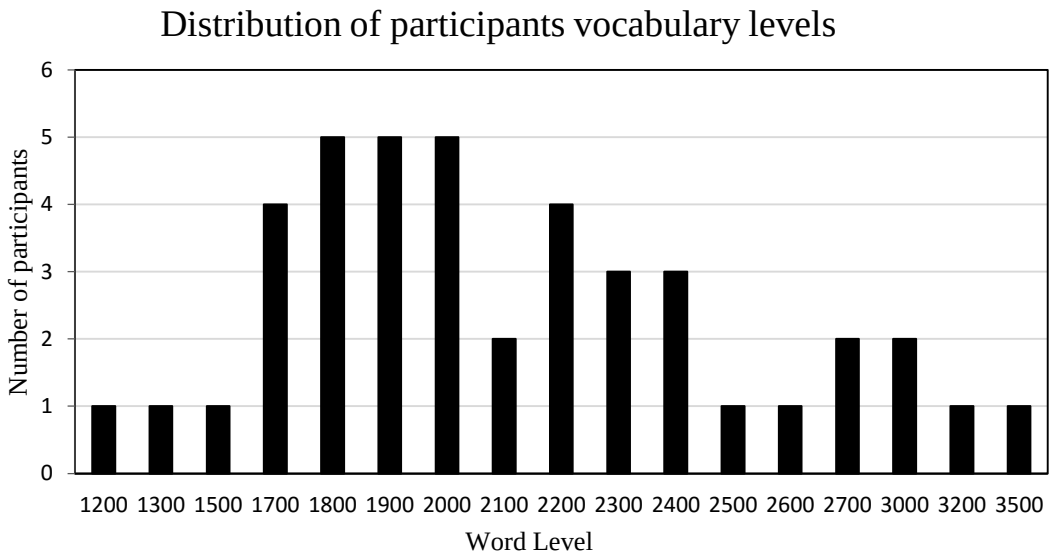


Figure 3.5 Distribution of estimated vocabulary levels of the participants

Interestingly, fifty percent of the participants fall within the 1700 to 2100 vocabulary levels. The graph follows a rough positively skewed distribution with a dip of two participants at the 2100 level and another dip of one participant each at the 2600 and 2700 vocabulary level.

Next, the participants were ordered from those with the highest to those with the lowest vocabulary (Figure 3.6a). After this, the participants were divided into two large groups based on their gender (Figure 3.6b). Then, following the order of biggest to smallest vocabulary sizes, the participants in each group were divided up into sets of three (Figure 3.7). The participants in each of these three sets were then randomly sorted, one set after the other, into the three intervention groups (Figure 3.8). For example, the top set of three male participants that had reached vocabulary levels of 3500, 3000, and 3000, were shuffled into the second, first and third participant conditions respectively based on the random roll of a die. If a participant got a roll of one or two, they were assigned to the first participant group, a three or four meant assignment to the second participant group, and on a roll of five or six the participants were assigned to the third participant group. This assignment process is called stratified random sorting for the purposes of this study.

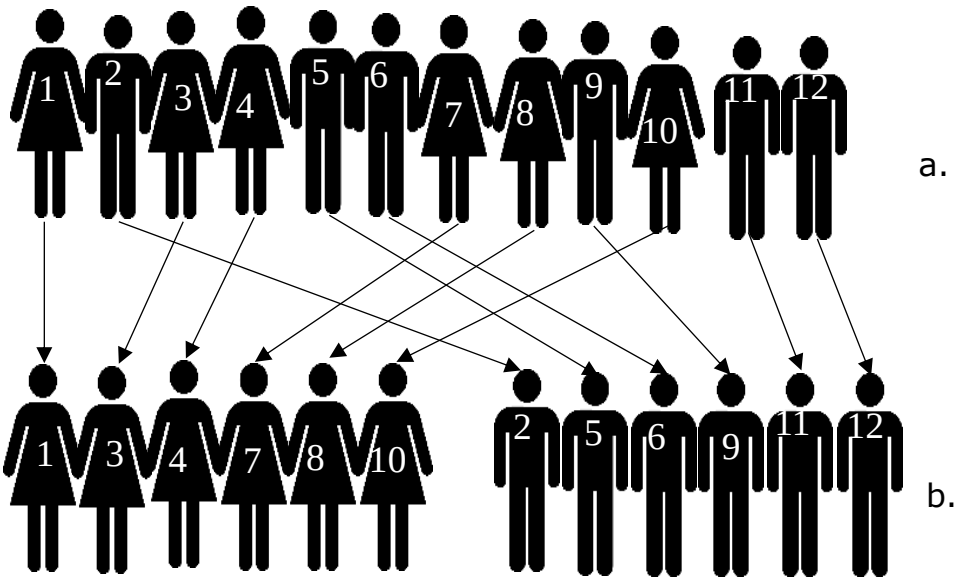


Figure 3.6 Participants ranked according to vocabulary levels (a) regrouped by gender (b).

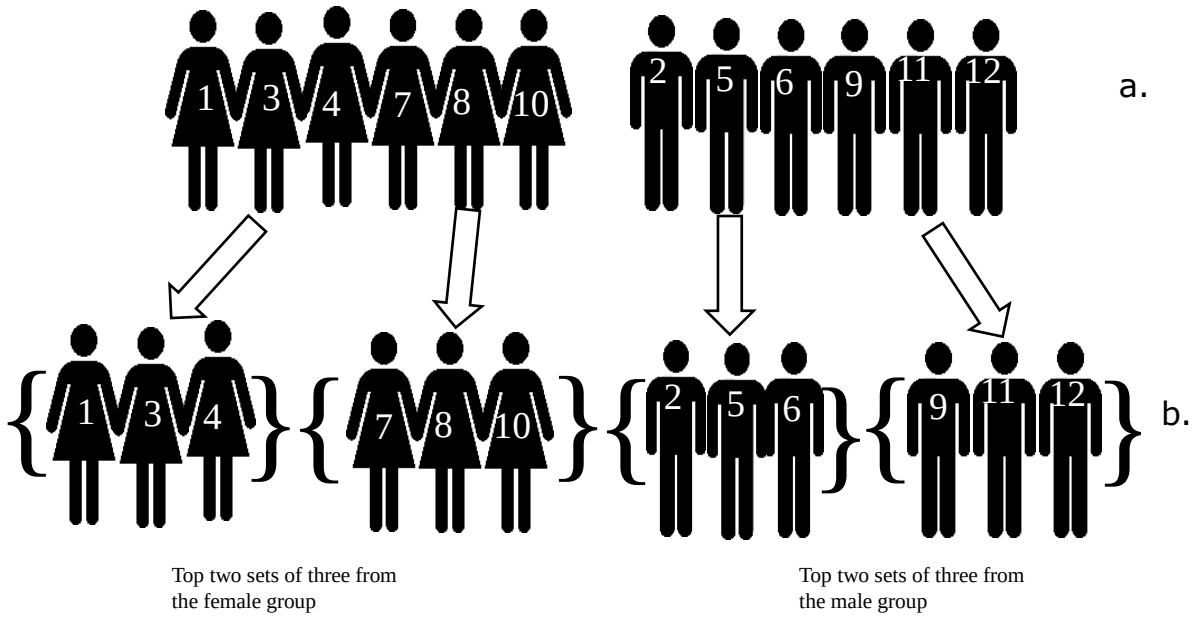


Figure 3.7 The participants in both groups (a) divided into sets of three (b).

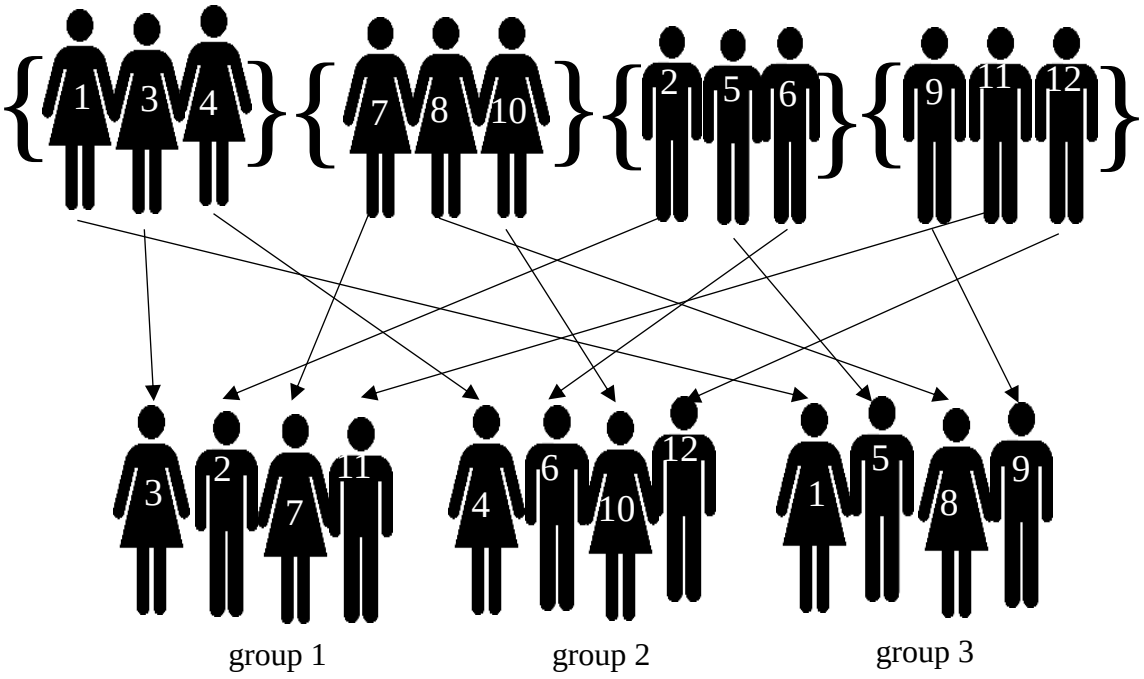


Figure 3.8. Participants randomly sorted three participant groups.

3.6.3 Vocabulary Learning Procedure

As far as the memorization of the target vocabulary is concerned, all the participants in this study used both types of experimental vocabulary learning conditions and they also acted as the control condition. To elaborate, each of the three intervention groups learnt seven words from the wordlist using the Combined Method, seven words from the wordlist using only Spaced Recall Practice, and seven words from the wordlist weren't studied at all. Specifically, group one learnt words one to seven using the Combined Method, words eight to 14 using Spaced Recall Practice, and didn't study words 15 to 21. Group two learnt words eight to 14 using the Combined Method, words 15 to 21 using Spaced Recall Practice, and didn't study

words one to seven. Group three learnt words 15 to 21 using the Combined Method, words one to seven using Spaced Recall Practice, and didn't study words eight to 14 (see Table 3.2).

Table 3.2. *Words memorized by participant groups.*

	Vocabulary words 1-7	Vocabulary words 8-14	Vocabulary words 15-21
Group 1 (students 1-14)	Combined Method	Spaced Recall	Control
Group 2 (students 15-28)	Control	Combined Method	Spaced Recall
Group 3 (students 29-42)	Spaced Recall	Control	Combined Method

Once the vocabulary learning phase of the treatment began, the students were given the following instructions. First, when the participants saw the LP's, they needed to use rote repetition to memorize word pairs without a keyword associated with them and used the keyword method they had learnt to memorize the word pairs that had a keyword and a descriptive sentence associated with them. They got 8 seconds for this procedure. This is the *learning procedure.* Second, when the participants saw the RP's, they needed to write down the English equivalent of the Chinese word they saw on the page. For this they received six seconds. This is the *recall procedure*. Third, they were not allowed to open the booklet until instructed to do so. The same instructions were written in Chinese on the cover of the booklets.

The first Vocabulary Learning Booklet were then given to the participants. The

participants were prompted by their instructor to read the instructions on the cover again. Following this the instructor played a recording that prompted the participants at the appropriate times to turn the pages in their booklets.

The first week's vocabulary booklet started with the Learning Phase. During the Learning Phase the learning procedure preceded the recall procedure. The Learning Phase was then followed by two Recall Phases. During each Recall Phase the recall procedure preceded the learning procedure. Finally, for the following three weeks, on the same day and at the same time, the students received three more vocabulary learning booklets consisting of only three consecutive Recall Phases (Table. 3.3).

Table 3.3 *Instructional treatment procedure*

Week	Step	Procedure	Participants		
1.1	1	Introduce the Keyword Mnemonic method as described in section 3.2.3.1 to all the participants	All Participants N(42)		
1.2		Review and recall the Keyword Mnemonic method as described in section 3.2.3.1 to all the participants	All Participants N(42)		
1.3	2	Have participants write Nation's Vocabulary Levels test. Then based on the results divide the participants into three Participant Groups (PG)	Participants were divided into male and female groups and then ordered from highest to lowest into sets of three. All the participants in each set of three were then shuffled into Participant Group A (PG A), Participant Group B (PG B), and Participant Group C (PG C).		
2	3	Introduce the Vocabulary Learning method as described in section 3.2.3.2	Participant Group A(PG A)	Participant Group B(PG B)	Participant Group C(PG C)
2	4	Complete Vocabulary Learning Booklet: Part 1 Learning Phase*; Part 2 & Part 3 Recall Phases** (Word order differed in each Phase)	PG A: Study Words 1-7: Combined Method Study Words: 8-14: Spaced Recall	PG B: Study Words 8-14: Combined Method Study Words: 15-21: Spaced Recall	PG C: Study Words 15-21: Combined Method Study Words: 1-7: Spaced Recall
3	5	Complete Vocabulary Retrieval Booklet (same as Vocabulary Learning Booklet except no Learning Phase and three Recall Phases, also word order differed in each Phase)	PG A: Retrieve Words 1-7: Combined Method Retrieve Words: 8-14: Spaced Recall	PG B: Retrieve Words 8-14: Combined Method Retrieve Words: 15-21: Spaced Recall	PG C: Retrieve Words 15-21: Combined Method Retrieve Words: 1-7: Spaced Recall
4	6	Complete Vocabulary Retrieval Booklet (same as previous Vocabulary Retrieval Booklet except word order differed in each Phase)	PG A: Retrieve Words 1-7: Combined Method Retrieve Words: 8-14: Spaced Recall	PG B: Retrieve Words 8-14: Combined Method Retrieve Words: 15-21: Spaced Recall	PG C: Retrieve Words 15-21: Combined Method Retrieve Words: 1-7: Spaced Recall
5	7	Complete Vocabulary Retrieval Booklet (same as previous Vocabulary Retrieval Booklet except word order differed in each Phase)	PG A: Retrieve Words 1-7: Combined Method Retrieve Words: 8-14: Spaced Recall	PG B: Retrieve Words 8-14: Combined Method Retrieve Words: 15-21: Spaced Recall	PG C: Retrieve Words 15-21: Combined Method Retrieve Words: 1-7: Spaced Recall
5	8	Redistribute the participants from the three participant groups into the three posttest groups	A third of each Participant Group was redistributed into each of the three Posttest Groups so that each posttest group would have roughly the same amount of participants from each of the three Participant Groups.		
5	9	**Immediate Posttest Group N(12)** Complete immediate posttest Written immediately after completion of treatment	$\frac{1}{3}$ of PG A: Scoring of Test Words 1-7: Combined method Words 8-14: Spaced Recall Words:15-21: Control	$\frac{1}{3}$ of PG B: Scoring of Test Words 1-7: Control Words 8-14: Combined method Words:15-21: Spaced Recall	$\frac{1}{3}$ of PG C: Scoring of Test Words 1-7: Spaced Retrieval Words 8-14: Control Words:15-21: Combined method
6	10	**First Delayed Posttest Group N(13)** Complete first delayed posttest Written one week after completion of treatment	$\frac{1}{3}$ of PG A: Scoring of Test Words 1-7: Combined method Words 8-14: Spaced Recall Words:15-21: Control	$\frac{1}{3}$ of PG B: Scoring of Test Words 1-7: Control Words 8-14: Combined method Words:15-21: Spaced Recall	$\frac{1}{3}$ of PG C: Scoring of Test Words 1-7: Spaced Retrieval Words 8-14: Control Words:15-21: Combined method
9	11	**Second Delayed Posttest Group N(17)** Complete second delayed posttest Written four weeks after completion of treatment	$\frac{1}{3}$ of PG A: Scoring of Test Words 1-7: Combined method Words 8-14: Spaced Recall Words:15-21: Control	$\frac{1}{3}$ of PG B: Scoring of Test Words 1-7: Control Words 8-14: Combined method Words:15-21: Spaced Recall	$\frac{1}{3}$ of PG C: Scoring of Test Words 1-7: Spaced Retrieval Words 8-14: Control Words:15-21: Combined method

3.7 Data collection procedure

After the participant groups completed the treatment procedure, they had to be once again divided into the three posttest groups. A stratified random sorting method was used for a second time to sort each of the participant groups into an immediate posttest group, a one week delayed posttest group (the first delayed posttest), and a four week delayed posttest group (the second delayed posttest) respectively (Figure 3.9).

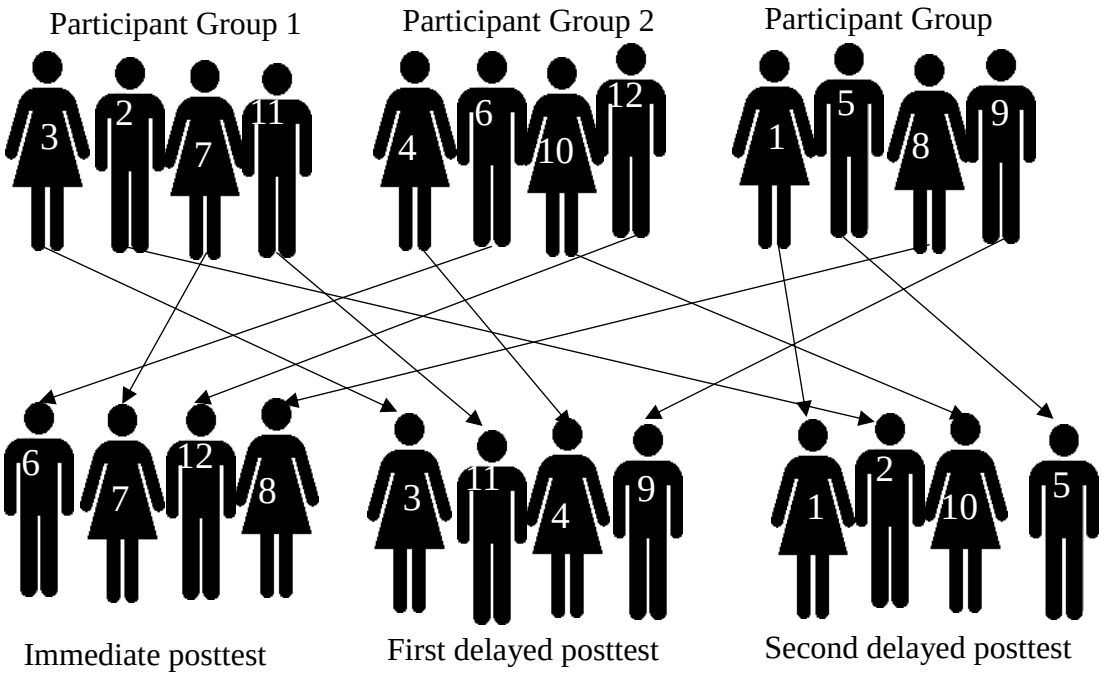


Figure 3.9. Participants randomly sorted into three posttest groups.

Each of the three posttest groups (immediate, first delayed, and second delayed) only did one of the posttests each. This was done in order to prevent the testing effect from becoming a confounding effect. If the students were to have taken more than one

of the posttests, the preceding posttests would have influenced the results of the following posttests. This is because taking the posttests themselves would increase the likelihood of successful recall on any subsequent posttests (Roediger & Karpicke; 2006b), thus confounding the results. The immediate posttest group took the posttest immediately after completing the treatment. The first delayed posttest group took the posttest one week after completing the treatment, and the second delayed posttest group took the posttest four weeks after the completion of the treatment.

Each participant received the Vocabulary Retention Test sheet at the start of the class on the day of their posttest group's scheduled test. Next, the instructor explained that they had to write down the English equivalent of all the Mandarin cue words they saw. Next to the prompts were lines where the participants could write down the L2 English target words. An answer was considered correct if a participant wrote down the correct L2 English target word for its L1 Chinese prompt. Minor spelling errors that did not affect the understanding of the words were allowed because the spirit of the Keyword Mnemonic in the Combined Method condition is essentially aural (Atkinson & Raugh, 1975; Raugh & Atkinson, 1975). For example, if a participant wrote *Mermade* instead of *Mermaid* the answer would be considered correct. In order to control for rater bias for what constitutes an understandable misspelled word, all misspelled words were given to 10 native English speaker ESL teachers to assess. The

words that were unanimously assessed as being understandable were accepted. The participants were given as much time as they need to finish the test. The participants were also instructed to hand in their tests as soon as they had finished the tests. The class continued with the regular curriculum after the final student had handed in their test paper.

Because each posttest group consisted of a third of each participant group and because they used different methods to study the word lists, special care was needed to correctly assign the scores for each treatment condition of each posttest. The same procedure was followed to assign the posttest scores for the immediate, first delayed and second delayed posttest. For Participant Group A words 1 to 7 counted to the score of the Combined Method condition, words 8 to 14 counted to the score of the Spaced Recall Practice condition, and words 15 to 21 counted to the score of the Control condition. For Participant Group B words 8 to 14 counted to the score of the Combined Method condition, words 15 to 21 counted to the score of the Spaced Recall Practice condition, and words 1 to 7 counted to the score of the Control condition. For Participant Group C words 15 to 21 counted to the score of the Combined Method condition, words 1 to 7 counted to the score of the Spaced Recall Practice condition, and words 8 to 15 counted to the score of the Control condition (see Appendix D).

3.8 Data analysis procedure

To see if the three conditions (the Combined Method condition, Spaced Recall Practice condition, and Control condition) had a significantly different effect on the amount of words retained a repeated-measures ANOVA was run for the immediate posttest, the one week delayed posttest and the one month delayed posttest. The reason repeated-measures ANOVAs were used is that the same participants had been exposed to the different learning conditions. Because this study is only interested in discovering if the Combined Method condition is significantly better than the Spaced Recall Practice condition, one tailed t-tests were used in the LSD post hoc analysis. The reason for this is if there is no significant difference between the two conditions, then the additional effort involved in producing keywords for the Combined Method condition would make the Spaced Recall Practice condition the preferable choice in a pedagogical context.

A one way analysis of the variances was used to determine if there were any significant differences between the immediate posttest, the first delayed posttest and the second delayed posttest. If the null hypothesis had not been rejected, it would have indicate that the L2 vocabulary that the students were exposed to were already known to them and the results of the study would have been invalid. This did not happen.

Furthermore, it also indicated where the rate of forgetting was significant. This helped give additional information about the shape of the forgetting curves.

Table 3.4 *Data analysis of the three research questions*.

Research questions	type of analysis
Do the participants retain significantly more L2 English words on the immediate posttest when using the Combined Method than when using Spaced Recall Practice?	Repeated measures ANOVA with one tailed LSD post hoc analysis
Do the participants retain significantly more L2 English words on the first delayed posttest when using the Combined Method than when using Spaced Recall Practice?	Repeated measures ANOVA with one tailed LSD post hoc analysis
Do the participants retain significantly more L2 English words on the second delayed posttest when using the Combined Method than when using Spaced Recall Practice?	Repeated measures ANOVA with one tailed LSD post hoc analysis
Are there any significant differences between the immediate posttests, the one week delayed posttests, and the two week delayed posttests?	One way ANOVA with a Scheffe post hoc analysis

3.9 Pilot study

A pilot study was performed in 2015, starting on the 6th of November and concluding on December 5th. One class of 18 students, who did not take part in the actual study, was recruited for the pilot study. The participants received one week of training on how to use the Mnemonic method as described in section 3.2.5.1 and then started the treatment on the 12th of November. The treatment concluded on the 19th of November.

3.9.1 Pilot study's Procedure

Before the treatment started the 18 students were sorted into three counter balance participant groups of 6 students each according to the stratified random sorting method described in section 3.6.2. Group One studied vocabulary words 1 to 7 using the Combined Method condition, vocabulary words 8 to 14 using the Spaced Recall Practice condition, and didn't study vocabulary words 15 to 21 at all. Group Two studied vocabulary words 1 to 7 using the Spaced Recall Practice condition, vocabulary words 15 to 21 using the Combined Method condition, and didn't study vocabulary words 8 to 14 at all. Group Three studied vocabulary words 8 to 14 using the Combined Method condition, vocabulary words 15 to 21 using Spaced Recall Practice condition, and didn't study vocabulary words 1 to 7 at all. On the 12th of November all three groups undergone the first treatment session using the Week One booklet following the procedures described in section 3.2.3.1. On the 19th of November, all the groups underwent the second and final treatment session using the Week Two booklet following the procedures described in section 3.2.3.1. The pilot differs from the actual study in that only two treatment sessions were undertaken, whereas in the actual study four treatments sessions were undertaken, one per week for four consecutive weeks. Also, at the conclusion of the final treatment session, instead of rearranging the groups into three posttest groups, they were only sorted into

two posttest groups. This was done in order to increase the number of participants per group in order to increase the strength of the statistical analysis. The group that did the immediate posttest also did the second delayed posttest. The other group only did the first delayed posttest. The immediate posttest was given directly after the conclusion of the final treatment session. The first delayed posttest was given one week after the conclusion of the final treatment session, and the second delayed posttest was given two weeks after the conclusion of the final treatment session. One further difference between the pilot study and the actual study is the retention intervals of the second delayed posttest. In the pilot study the RI of the final posttest was two weeks, whereas the RI of the actual study was four weeks.

Table 3.5.*Results of the posttests*.

	control		spaced		combined	
	mean	SD	mean	SD	mean	SD
immediate posttest	0.56	1.01	6.67	0.71	6.78	0.44
first delayed posttest	0.89	0.78	4.12	1.90	4.34	2.12
second delayed posttest	0.56	1.01	4.78	1.72	5.78	1.39

3.9.2 Results

First, the results for the immediate posttest showed that the mean score was 6.78 (N=9) with a standard deviation of 0.44 for the Combined Method condition and 6.67 (N=9) with a standard deviation of 0.71 for the Spaced Recall Practice condition. The results of a Pair sampled T-test revealed no significant difference between in the

immediate posttest between the Combined Method condition and the Spaced Recall Practice condition ($p=0.35$). Second, for the first delayed posttest, the mean score for the Combined Method condition was 4.34 ($N=9$) with a standard deviation of 2.12 and for the Spaced Recall Practice condition was 4.12 ($N=9$) with a standard deviation of 1.90. Again a paired sample T-test revealed no significant difference between the Combined Method condition and the Spaced Recall Practice condition ($p=0.71$). Finally, In a two way ANOVA testing for effects of the Retention Interval, intervention condition, and interactions between the two found significant effects for the RI on the number of words retained, $F(1, 27)= 18.82$, $p<0.001$, and for the intervention conditions on the number of words retained, $F(2, 18)=76.24$, $p<0.001$.

On closer analysis using repeated paired sample T-tests, a significant difference was found between the immediate posttest control condition and the Combined Method condition ($p<0.001$), and a significant difference was found between the immediate posttest control condition and the Spaced Recall Practice condition ($p<0.001$). Also, a One-way ANOVA showed a significant effect of the RI of the Combined conditions on the vocabulary retention, $F(1,16) = 11.46$, $p=0.004$, and of the RI of the Spaced Recall Practice on the vocabulary retention, $F(1, 16)= 14.30$, $p=0.002$. No effect of the RI of the control groups on vocabulary retention were found, $F(1, 16)= 0.61$, $p=0.45$.

The second delayed posttest results seemed problematic as they yielded higher mean scores than the first delayed posttest. This might have been a direct result of using the participants that wrote the immediate posttest to write the second delayed posttest. The mean of the Combined Method condition was 5.78 (N=9) with a standard deviation of 1.39 and the mean of the Spaced Recall Practice condition was 4.78 (N=9) with a standard deviation of 1.72. A one-tailed pair samples T-test comparing the Combined Method with the Spaced Recall Practice condition showed a marginally nonsignificant difference (p=0.054).

3.9.3 Item analysis

In order to see if any of the items on the Vocabulary Retention Tests could have skewed the results and item analysis was run using the immediate posttest results as the R_u and the first delayed posttest results as the R_l. The results can be seen in table 6. Four items in the analysis had a discrimination index of 0 or less. The Items were sponge (0), crosswalk (0), ranch (-0.11), and doorway (-0.11). These items may have a higher frequency in modern English than originally estimated, or their frequency might be higher in the media that the participants consume. An example of this might be sponge, which has a vocabulary frequency of 9489 on the corpus of contemporary English's frequency list. However, because of programs like Sponge Bob square pants, the current generation of middle schoolers may have had many more encounters with

this word.

Table3.6 *Item analysis of the Productive Vocabulary test*.

Item number	difficulty index	discrimination index
1	0.5	0.11
2	0.56	0.22
3	0.56	0.44
4	0.56	0.44
5	0.56	0.22
6	0.61	0.11
7	0.56	0.22
8	0.39	0.33
9	0.39	0.56
10	0.61	0.11
11	0.78	**0**
12	0.5	0.33
13	0.67	**0**
14	0.39	0.56
15	0.45	0.44
16	0.61	0.11
17	0.83	**-0.11**
18	0.5	0.33
19	0.39	0.11
20	0.72	**-0.11**
21	0.5	0.33

Items 11, 13, 17, and 20 were found to have discrimination levels of 0 or below, indicating that they were already known to the students or just too easy to be included in the study.

3.9.4 Changes made to the actual study

The first change that needed to be made to the actual study was the abandonment of using any of the test groups more than once. Having the students

write the immediate posttest and the second delayed posttest had a profound testing effect on the participants as was seen in the results of the second delayed posttest.

The next change was to replace the items that scored 0 or less on the discrimination index of the item analysis with new items. In order to do this, the next 20 words on the randomized filtered GEPT high intermediate wordlist were used in a small scale memorization test in the class of participants that took part in the pilot study. The students first wrote a pretest on the words. Then, after completing a memorization intervention and a one day delayed posttest, another item analysis was performed, and items that score in the mid-levels of the difficulty index and discrimination index were chosen to replace the non-performing items. The participants themselves helped with creating the keywords for the new items after the item analysis had been completed.

The final changes made to the actual study was the timing of the second delayed posttest and the duration of the intervention period. Because a marginal nonsignificant difference was only found between the Combined Method condition and the Spaced Recall Practice condition on the second delayed posttest, any differences in the treatment effects may only have become apparent after a longer RI. This may be due to the shape of the forgetting curves that may only vary slightly, but as time passes, the slight changes may become more exaggerated. Also, the intervention period was

extended from two weeks to four weeks. This was done to increase the chances that the initial benefit seen in productive performance of the combined strategy in the study of Fritz et al (2007) would be retained.

CHAPTER FOUR

RESULTS AND DISCUSSIONS

This chapter will describe the data collected by the vocabulary retention test administered at three time points (immediate posttest, one-week delayed posttest, and four-week delayed posttest). The descriptive statistics of the participants' performances on the three posttests will first be presented followed by the inferential statistics in the order of the four research questions stated in Chapter One. Finally any unexpected results will be discussed and then a brief summary of the chapter will be given at the end of this chapter.

4.1 Results of the immediate posttest

The immediate posttest consisted of a paper on which the list of 21 L1 Chinese prompts was printed. Next to the prompts were lines where the participants could write down the L2 English target words. An answer was considered correct if a participant wrote down the correct L2 English target word for its L1 Chinese prompt. As is stated in section 3.7 minor spelling errors were acceptable if they didn't influence the understanding of the word.

The test had three unmarked sections, the Combined Method words, the Spaced Recall Practice words, and the Control words (see Appendices C and D). These sections were unmarked on the test paper because each participant group studied each

of the word sections under different conditions The immediate posttest was administered directly after the participants completed the final vocabulary retrieval booklet. Sixteen of the original forty-eight participants were selected using the stratified random sorting method to complete the first posttest. During the stratified random sorting process the students are grouped hierarchically from highest to lowest in sets of three according to the results of their Vocabulary Levels Test scores. The participants in each set are then randomly shuffled into one of the three participant groups. However, due to attrition, only twelve students completed the test. Because four participants were required to take an extracurricular activity, they were not in the class during the time the immediate posttest was administered.

4.1.1 Descriptive statistics of the immediate posttest

Table 4.1 shows the results (to a maximum of seven) of the three vocabulary sections of the immediate posttest: the Combined Method, Spaced Recall Practice, and Control condition. The first column contains the identification code for each student. The participants each received a one or two letter identification code for example A to Z and then AA to AP (see section 3.1). The second, third and fourth columns show the participants' test performance on the Combined Method, the Spaced Recall Practice and Control condition sections of the immediate posttest. Each section contained seven words and each word correctly recalled counted one point to

the score of the immediate posttest performance on that section. With the exception of Participants A and B, all other participants that wrote the first posttest got all the words correct for both the Combined Method and the Spaced Recall condition. Participant A only got two of the Combined Method words correct and Participant B got six of the words correct. Regarding the Spaced Recall condition words, Participant A got four words correct and Participant B got five correct. Moving on to the control condition words, Participant C got one of the words correct. All other participants couldn't retrieve any of the control words.

Table 4.1 *Results of the immediate posttest per student*

Participant (N=12)	Combined Method	Spaced Recall	Control
A	2	4	0
B	6	5	0
C	7	7	1
D	7	7	0
E	7	7	0
F	7	7	0
G	7	7	0
H	7	7	0
I	7	7	0
J	7	7	0
K	7	7	0
L	7	7	0

Figure 4.1 shows the distribution of the number of participants over the number of vocabulary words (to a maximum of seven per section) recalled. These results reflect Table 4.2 closely. In both the Combined Method and the Spaced Recall

condition a ceiling effect can be observed. This shows that, compared to the Control, both the Combined Method and the Spaced Recall condition were equally effective. Of further note, only one participant successfully retrieved a Control condition word. The word in question was *vein*. When enquired about how the word was learned or when, the participant couldn't remember. However, this result was an outlier and statistically insignificant. In actual fact, no other Control condition words were successfully retrieved. This clearly indicates that the target vocabulary words were significantly unknown to the participants.

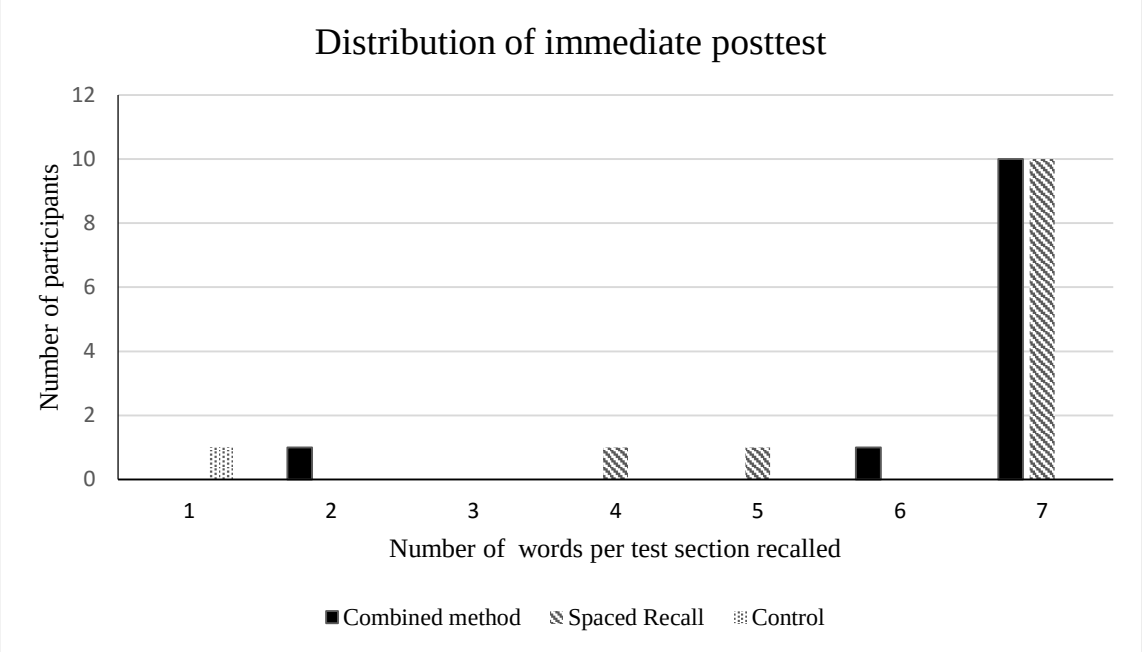


Figure 4.1 Distribution of immediate posttest

Next, means and standard deviations of the immediate posttest scores were calculated for each learning condition. The results are shown in Table 4.3. The mean of the Combined Method was 6.500 and the standard deviation was1.446. The mean of the Spacer Retrieval condition was 6.583 and the standard deviation was 0.996.

The mean of the Control condition was 0.083 and the standard deviation was 0.289.

Table 4.2 *Means & Standard deviations the immediate posttest*.

	Mean	SD
combined method	6.500	1.446
Spaced Recall	6.583	0.996
control	0.083	0.289

4.1.2 Inferential statistics of the immediate posttest

To see if the three conditions (the Combined Method condition, Spaced Recall condition, and Control condition) had a significantly different effect on the amount of words retained a repeated-measures ANOVA was run. Because the assumption of sphericity was rejected (Mauchly, 1940), the Greenhouse-Geisser correction to the degrees of freedom was used. The repeated-measures ANOVA revealed a significant difference between the number of words retained by using the Combined Method, Spaced Recall Practice, and the Control condition $F(1.2, 13.196) = 283.592$, $p < 0.001$.

An LSD post hoc analysis was run to determine which pairwise comparisons were significant. The post hoc analysis revealed no significant difference between the Combined Method and Spaced Recall Practice in a one tailed comparison ($MD = 0.083$; $p = 0.337$). However, significant differences did exist between the one tailed comparisons of the Control condition and the Combined Method condition ($MD=5.417$; $p<0.001$) and the Control condition and the Spaced Recall Practice condition ($MD = 6.50$; $p <0.001$). In summary, no significant difference was found

between the Combined Method condition and the Spaced Recall condition, but significant differences were found between the Control condition and both the Combined Method and the Spaced Recall conditions.

4.2 The first delayed posttest

The first delayed posttest was exactly the same as the immediate posttest except for one aspect. It was administered one week after the participants completed of the fourth (and final) vocabulary memorization booklet (see section 3.2.2). Sixteen participants that hadn't written the first posttest were selected using the stratified random sorting method (briefly described in section 4.1.1) to complete the first delayed posttest. Due to attrition, only 13 complete the first delayed posttest. Because one participant was absent due to illness and two were required to attend a science competition, three students were absent from the class when the first delayed posttest was administered.

4.2.1 Descriptive statistics of the first delayed posttest

Table 4.4 shows the results (to a maximum of seven) of the three vocabulary sections of the first delayed posttest: the Combined Method, Spaced Recall Practice, and Control condition. The first column contains the identification code for each student. The second, third and fourth columns show the participants' test performance on the Combined Method, the Spaced Recall Practice and Control condition sections

of the first delayed posttest. As already stated above, each word correctly recalled counted one point to the score of the first delayed posttest performance on that section. Of note in the results of the first delayed posttest, no participants recalled any of the control words. Furthermore, Participants X & Y recalled all the words from both the Combined Method condition and the Spaced Recall condition. Also, Participants V and W only recalled all the words from the Spaced Recall condition while Participant N only recalled all the words from the Combined Method condition. Participant M had the lowest score, recalling three words from the Combined Method condition and three words from the Spaced Recall condition.

Table 4.3 *Results of the first delayed posttest*

Participants (N=13)	Combined Method	Spaced Recall	Control
M	3	3	0
N	7	4	0
O	5	5	0
P	5	5	0
Q	6	5	0
R	5	6	0
S	6	6	0
T	6	6	0
U	6	6	0
V	6	7	0
W	6	7	0
X	7	7	0
Y	7	7	0

Figure 4.3 shows the distribution of the number of participants over the number of vocabulary words (to a maximum of seven per section) recalled for the first

delayed posttest (Figure 4.3). It is apparent that the results are slightly more spread out than that of the immediate posttest. Although a celling effect isn't clearly visible in this distribution, the students are clustered around the high end of the graph. Of the participants, 46% recalled six Combined Method condition words which is also the mode. Furthermore, only 31% of the participants recalled less than six Combined Method condition words. Almost 62% of the participants correctly recalled six or more Spaced Recall condition words. Also, only 15% of the participants correctly recalled four or less Spaced Recall condition words. Additionally, no participants recalled less than three words from the either the Combined Method condition or the Spaced Recall condition. Finally, the mode of the Combined Condition was six and for the Spaced Recall condition six and seven. The clustering of the results near the high end or the range indicates that a mild ceiling effect may indeed be present in the first delayed posttest, although not as obviously observable as in the immediate posttest.

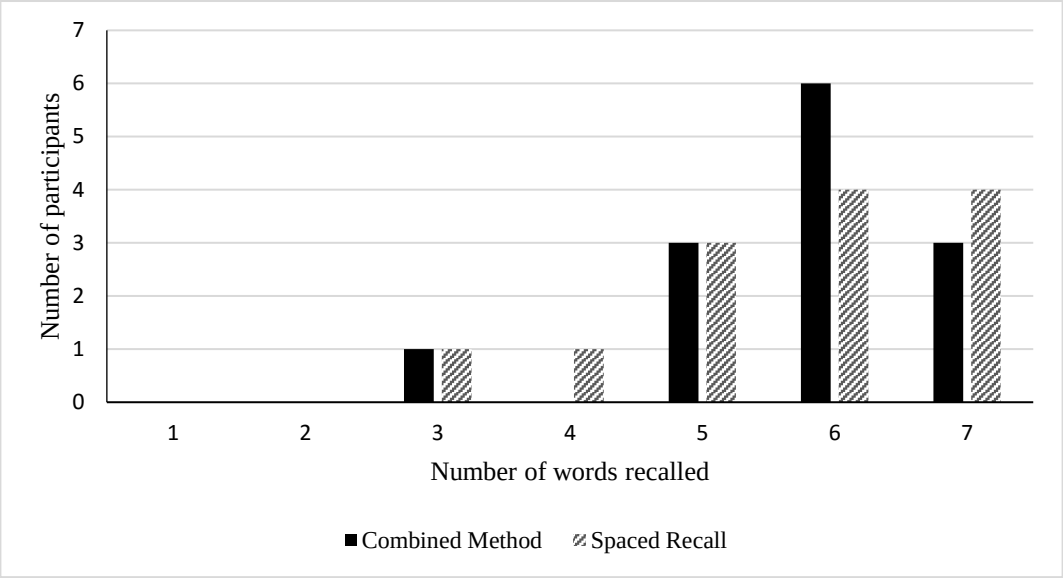


Figure 4.2 Distribution of first delayed posttest

Moving on, the means and the standard deviations of the first delayed posttest were calculated for the three learning conditions. The results are shown in Table 4.4. The mean of the Combined Method was 5.769 and the standard deviation was 1.251. The mean of the Spaced Recall condition was 5.692 and the standard deviation was 1.092. Because all the participants who wrote the first delayed posttest failed to recall any of the Control condition words, both the mean and the standard deviation are 0.

Table 4.4 *Means & Standard deviation of the first delayed posttest.*

	Mean	SD
combined method	5.769	1.092
Spaced Recall	5.692	1.251
control	0	0

4.2.2 Inferential statistics of the first delayed posttest

To see if the amount of words retained by the participants were affected differently by the Combined Method, Spaced Recall , or Control conditions a repeated-measures ANOVA was run. The assumption of sphericity was not rejected (Mauchly, 1940), thus no corrections to the degrees of freedom were made. The repeated-measures ANOVA revealed a significantly different effect of the Combined Method, Spaced Recall Practice, and the Control conditions on the number of words retained $F(2, 24) = 22.783$, $p < 0.001$.

An LSD post hoc analysis was run to determine which pairwise comparisons were significant. The post hoc analysis revealed no significant difference between the Combined Method condition and the Spaced Recall Practice condition results of the first delayed posttest in a one tailed comparison ($MD = 0.077$; $p = 0.397$). However, significant differences did exist between the one tailed comparisons of the Control condition and the Combined Method condition ($MD=5.769$; $p<0.001$) and the Control condition and the Spaced Recall Practice condition ($MD = 5.692$; $p < 0.001$) results. In sum, no significant difference was found between the Combined Method condition and the Spaced Recall condition results of the first delayed posttest, but significant differences were found between the Control condition and both the Combined Method and the Spaced Recall condition results.

4.3 The second delayed posttest

The second delayed posttest was exactly the same as the immediate posttest. It was administered four weeks after the completion of the fourth (and final) vocabulary memorization booklet was completed. 17 participants that had not completed either the first or the first delayed posttests were selected to complete the final posttest. All of the original 17 participants completed the posttest.

4.3.1 Descriptive statistics of the second delayed posttest

The results (to a maximum of seven) of the three vocabulary sections of the second delayed posttest (the Combined Method, Spaced Recall Practice, and Control condition) are shown in Table 4.6. The first column contains the identification code for each student. The second, third and fourth columns show the participants' test performance on the Combined Method, the Spaced Recall Practice and Control condition sections of the second delayed posttest respectively. Of note, none of the participants recalled any of the Control condition words. Also, only participant AP recalled all of the words from both the Combined Method condition and the Spaced Recall condition sections. Participant AL only recalled all the words from the Combined Method condition section and no participants recalled all of the words from the Spaced Recall condition. Four participants recalled more words from the Spaced Recall condition section than from the Combined Method condition section. Ten

students recalled more words for the Combined Method condition than for the Spaced Recall condition. Finally, three participants recalled the same number of words for both the Combined Method and the Spaced Recall conditions.

Table 4.5 *Results of the second delayed posttest*

Participants (N=15)	Combined Method	Spaced Recall	Control
Z	1	2	0
AA	3	2	0
AB	3	2	0
AC	6	2	0
AD	2	3	0
AE	4	3	0
AF	4	3	0
AG	5	3	0
AH	5	3	0
AI	4	4	0
AJ	5	4	0
AK	6	4	0
AL	7	4	0
AM	5	5	0
AN	4	6	0
AO	5	6	0
AP	7	7	0

Figure 4.3 shows the distribution of the number of students over the number of words recalled for the second delayed posttest. This graph differs from the previous graphs because the distribution is spread out over the entire range. Also, no ceiling effect is visible. Furthermore, the modes of the Combined Method condition results

and the Spaced Recall condition results for the second delayed posttest have a difference of two words whereas in the previous posttests no difference existed. The Combined Method condition. Additionally, the distribution of the Combined Method condition and the Spaced Recall condition appear to follow two separate rough normal distributions. This hints at different effects of the Combined Method condition and the Spaced Recall condition on the distributions of the second delayed posttest results.

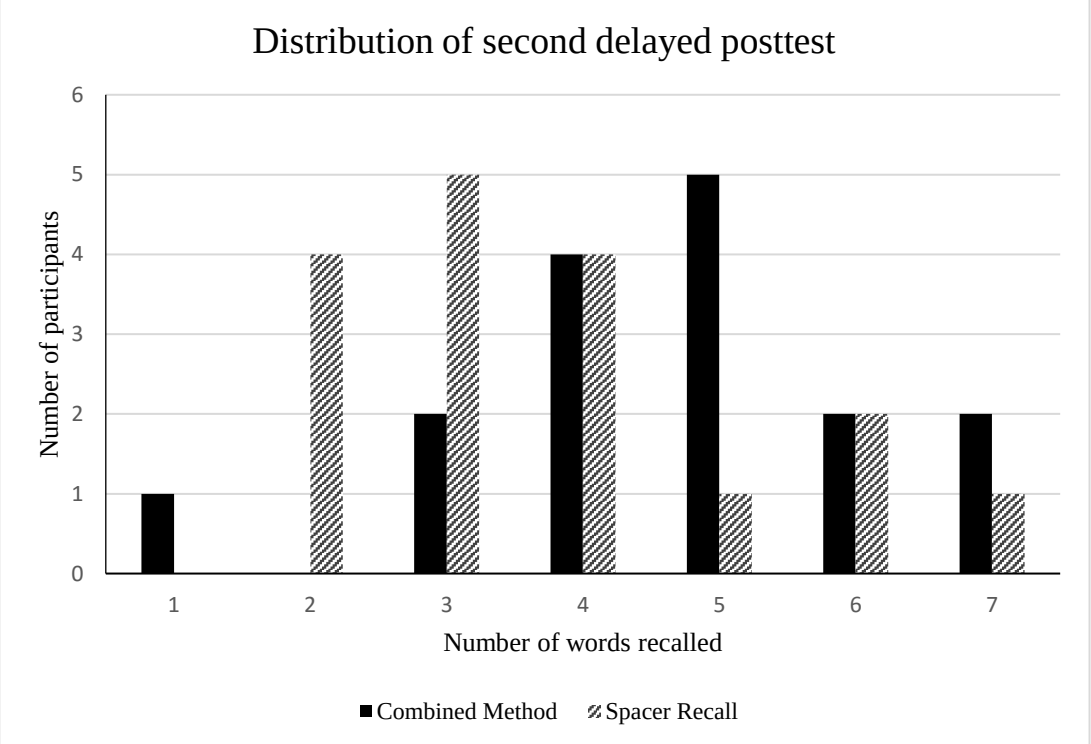


Figure 4.3 Distribution of second delayed posttest

Next, means and standard deviations of the second delayed posttest scores were calculated for each learning condition. The results are shown in Table 4.6. The mean of the Combined Method was 4.47 and standard deviation was 1.625. The mean of the

Spaced Recall condition was 3.706, and the standard deviation was 1.532. Because all the participants that wrote the second delayed posttest failed to recall any of the Control condition words, all the descriptive statistics for it were 0.

Table 4.6 *Mean and Standard Deviation second delayed posttest.*

	mean	std. dev.
control	0	0
Spaced Recall	3.706	1.532
combined method	4.471	1.625

4.3.2 Inferential statistics of the second delayed posttest

A repeated-measures ANOVA was run to investigate if the amount of words recalled on the second delayed posttest by the participants were affected differently by the Combined Method, Spaced Recall , or Control conditions. The conditions for teh assumption of sphericity was not met (Mauchly, 1940), thus a Greenhouse-Geisser correction to the degrees of freedom was made. The repeated-measures ANOVA revealed a significantly different effect of the Combined Method, Spaced Recall Practice, and the Control conditions on the number of words retained ($F(1.990, 39.844) = 78{,}527$, $p < 0.001$.)

Repeated one tailed pair-samples t-tests were conducted as a post hoc analysis to compare the means of the Combined Method condition and the Spaced Recall condition results, the Combined Method condition and the Control condition results, and the Spaced Recall condition and the Control condition results. First, a

significant difference between the Spaced Recall condition (M=3.706, SD=1.532) and the Combined Method condition results (M=4.471, SD=1.625); ($t_{(16)}$=2,018 p=0.031) was found. Because the analysis performed one-tailed t-tests, it can be inferred that the Combined Method condition was significantly better for the recall of the L2 words when prompted by L1 equivalents. Second, a significant difference was found between the Combined Method condition (M=4.471, SD=1.625) and the control condition (M=0.000, SD=0.000); $t(16)=11.345$, $p<0.001$ ($\alpha<0.05$), as well as between the Spaced Recall condition (M43.706, SD=1.532) and the control condition (M=0.000, SD=0.000); $t(16)=9.977$, $p<0.001$ ($\alpha<0.05$). This indicates two things: first, the words were not known to the participants before the study started, and second, that using either the Combined Method or the Spaced Recall conditions to memorize the words is better than doing nothing at all.

4.4 Comparing the means of the posttests longitudinally

In order to see if there were any significant forgetting that occurred between the first, second, and second delayed posttests of the Combined Method and the Spaced Recall condition respectively, a one-way ANOVA was run.

First, hypothesis testing was carried out to examine whether significant forgetting took place over the three time points for the Combined Method condition. There was a significant effect of the Combined Method posttests on words correctly

recalled for the three times ($F(2, 39) = 7.566$, $p= 0.002$). A post hoc comparisons using the Scheffe test indicated the mean score for the second delayed posttest (mean = 4.471, SD = 1.625) was significantly different ($p= 0.002$) to from the immediate posttest (mean = 6.500, SD = 1.446) but not significantly different ($p= 0.059$) from the first delayed posttest (mean= 5.769, SD= 1.092). Furthermore, the immediate posttest was not significantly different ($p= 0.449$) from the first delayed posttest. From these results it can be ascertained that a significant amount of forgetting of the Combined Method condition words did indeed take place between the immediate and the second delayed posttest. However, not much forgetting took place in the first week after the completion of the final vocabulary learning booklet. A marginally insignificant amount of words were forgotten between the second week and the fourth week after completion of the intervention.

Second, another hypothesis test was carried out to examine if significant forgetting took place between the results of the Spaced Recall condition sections of the three posttest. There was a significant effect for the Spaced Recall posttests of words correctly recalled for the three times. ($F(2, 39) = 18.546$, $p< 0.001$). A post hoc comparisons using the Scheffe test indicated that the mean score for the second delayed posttest (M = 3.706, SD = 1.532) was significantly different ($p< 0.001$) from the immediate posttest (M = 6.583, SD = 0.996) and significantly different ($p= 0.001$)

from the first delayed posttest (M= 5.692, SD= 1.251). However, the immediate posttest was not significantly different (p= 0.250) from the first delayed posttest. These results of the ANOVA indicate that a significant amount of forgetting of the Spaced Recall condition words did indeed take place between the first and the second delayed posttest and between the second and the second delayed posttests.

4.5 Discussion of the results

The results of the study will be discussed in accordance to the research questions and with reference to the literature review. Additionally, possible reasons for the results will be discussed and compared to current theory.

4.5.1 Immediate posttest

The null hypothesis of the first research question was: There will be no significant difference in the amount of vocabulary retained between the Combined Method and the Spaced Recall condition in the immediate posttest. Fritz et al., (2007) reported that there was no significant difference between their Combined Strategy (similar to the Combined Method in the current study) and the Retrieval Practice (similar to the Spaced Recall condition in the current study) in their immediate posttest and their one week delayed posttest. Their study inspired the design of the current study. Thus, it was expected that the null hypothesis for the immediate posttest would not have been rejected. This is indeed what happened.

There might be multiple reasons why the null hypothesis wasn't rejected. The first being that the addition of the Keyword Mnemonic to Spaced Recall Practice in the creation of the Combined Method may simply not yield an additional advantage over that already given by Spaced Recall Practice alone as observed in the Fritz et al., (2007) study. In their study, Spaced Recall Practice was more effective than the Keyword Mnemonic for productive vocabulary performance.

However, an alternative explanation might be that the ceiling effect may have contributed to the result in this experiment. As has already been demonstrated repeatedly in the literature, Spaced Recall as a rote memorization method is very effective (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke, & Bauernschmidt, 2011; Karpicke, & Roediger, 2010; Logan & Balota, 2008), and any benefit gained from combining it with the Keyword Mnemonic may only become apparent under more demanding conditions. This may happen if either more words need to be memorized, or if more time has elapsed for forgetting to occur. Considering that ten of the twelve participants that wrote the immediate posttest recalled all the words they were exposed to, this explanation might be highly plausible.

4.5.2 First delayed posttest

The null hypothesis of the second research question was: There will be no significant difference in the amount of vocabulary retained between the Combined

Method condition and the Spaced Recall condition in the first delayed posttest. It was expected that there would be no significant differences between the two conditions at this stage in line with Fritz et al., (2007). Thus the results of the first delayed posttest came as no surprise.

The lack of a significant difference between the Combined Method and Spaced Recall Practice may be due to the cognitive demands of using the Keyword Mnemonic in the Combined Method. In Spaced Recall Practice, new sound patterns have to be mapped onto an existing schema (Jiang, 2000). However, when using the Keyword Mnemonic, the participants also have to make both an acoustic link and an imagery link in order to use it successfully. Consequently, given the limited capacity of working memory of individuals (Baddeley, 2001), the Keyword Mnemonic may be more susceptible to decay over time. In other words, if either the acoustic link or the imagery link isn't successfully retrieved, the L2 English target word may also not be successfully retrieved. Thus, as Wang et al. (1992) pointed out, although there might be an initial benefit to using the Keyword Mnemonic, this benefit may be lost over time.

Following on from the ceiling effect explanation of the results of the first posttest, another reason for not seeing a significant difference between the testing conditions may be that not enough time had passed for any difference between the

methods to become apparent. As far back as the memory studies done by Ebbinghaus (1885), it has been known that forgetting takes place on some sort of curve that initially falls quite fast and then over time flattens out. Later on, Wixted and Ebbesen (1991) showed that this forgetting curve is best described by simple power functions. If the forgetting of the two conditions take place on different forgetting curves, an initial insignificant difference may become quite significant over time.

4.5.3 Second delayed posttest

The null hypothesis of the third research question was: There will be no significant difference in the amount of vocabulary retained between the Spaced Recall condition and the Combined Method condition in the final delayed posttest. Because of the possibility that the Combined Method and Spaced Recall Practice were leading to the forgetting taking place on two different forgetting curves, it was with the second delayed posttest that the possibility of rejecting the null hypothesis was the strongest, and this is what was found in the current study.

A possible explanation for the significant result in this study was postulated by Fritz et al., (2007). Because their study found a marginally non-significant difference between the Retrieval Practice condition and the other conditions over time, Fritz et al., (2007) state that with a longer retention interval the marginally insignificant difference they found might have been exacerbated leading to a significant difference.

This line of argument may also hold true for the Combined Method condition. Thus, because the retention interval of the second delayed posttest was four weeks, any marginal differences observed in the first delayed posttest may have been exacerbated to the point that a significant difference was observable in the second delayed posttest. Furthermore, the reason why the initial benefit seen in productive performance of the combined strategy in the Fritz et al., (2007) study was retained in this study might be because of the extension the intervention period from one session to four spread out over a month.

4.5.4 Forgetting curves of treatment conditions

Figure 4.4 shows the data points of the means of the two experimental conditions as well as forgetting curves that intersect those data points. The forgetting curves were created using Excel to generate trend-lines that extended four weeks into the future. Simple power functions were used to calculate the curves as the literature has shown that forgetting usually takes place on such curves (Wixted & Ebbesen, 1991). The formulas of the trend lines were then generated using Excel and these formulas were then further manipulated in order to create a better fit with the three data points. What this graph clearly demonstrates is that although at first there might seem to be no significant difference between the two conditions, as time progresses, the two functions diverge. Thus, as time passes, the difference between the two

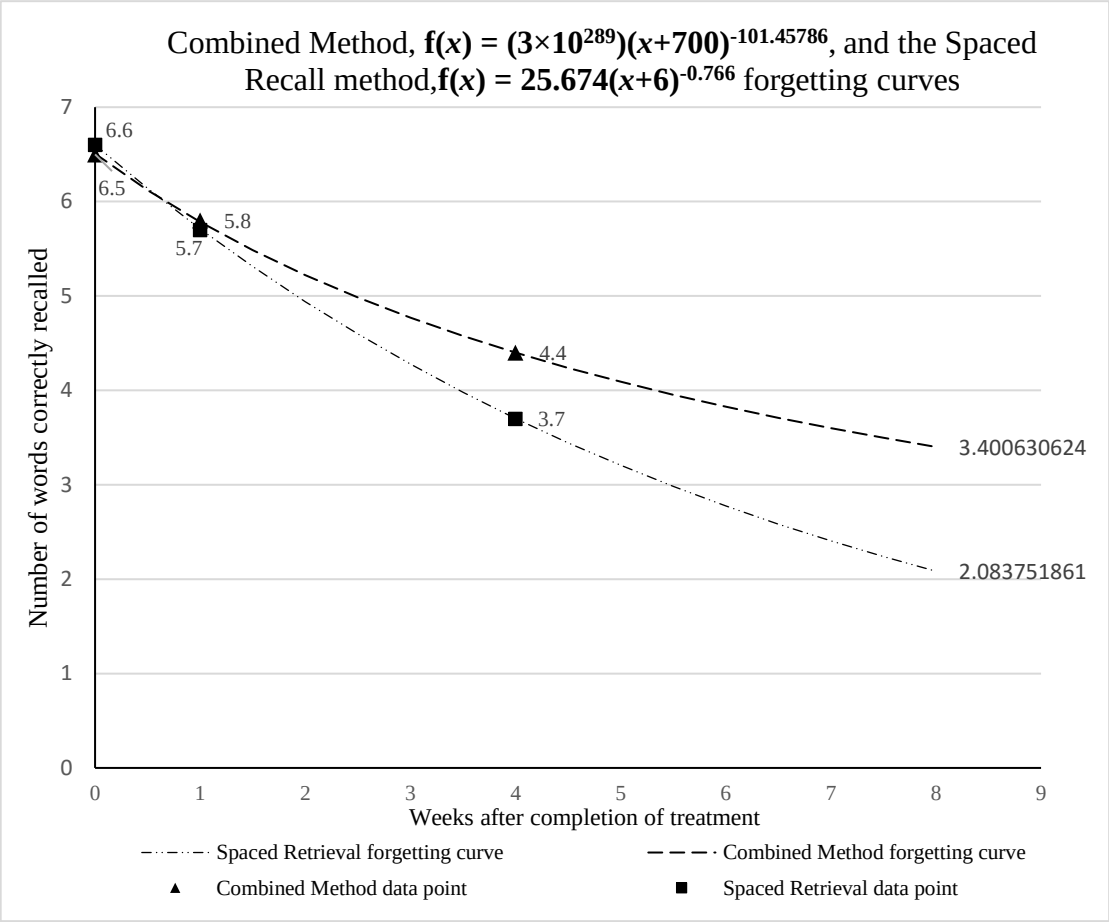


Figure 4.4 Forgetting curves of Combined Method and Spacer Retrieval Practice

conditions become more apparent, and the significant difference increases.

One possible theoretical explanation for the results may lie with Dual Coding Theory (Clark & Paivio, 1991). Dual coding theory proposes two separate mental stores for memories, one verbal and the other imagery. Thus by using the Keyword Mnemonic in the Combined Method and making both an acoustic link and an imagery link, both stores are accessed and thus the chances of successful retrieval are enhanced. If, for example, a student wants to recall the word “collection” from the

Chinese “搜藏” thinking of the image in the imagery store of a god figure with a cola bottle collection could activate “可樂神” in the verbal store, which in turn could activate the target L2 English word “collection” by direct connection between the two and also in the process reinforcing the aural component in the memory trace. A sort of spreading activation occurs. The drawback of this is that if the connection between “搜藏” and the associated image cannot be recalled or between the image and “可樂神”, then the spreading activation will not occur. This could explain why the Keyword Mnemonic alone seems to not produce durable results (Wang et al., 1992). However, Spaced Recall Practice may enhance all the memory traces by reminding the learner of a previous or the first instance of the memory trace, thus making retrieval possible and thus making the learner benefit from retrieval practice (Benjamin & Tullis, 2010; Dunlosky, et al, 2013). This in turn could make it easier for the learner to recall either the imagery link, the acoustic link, or the target L2 English word directly, further enhancing the chances of successful retrieval of the target L2 English word.

In conclusion, although at first the difference between Combined Method condition and the Keyword condition might seem insignificant, over time the difference might become more apparent as the memories created through the two conditions deteriorate along different forgetting curves. The reason the Combined Method seems to lead to statistically significantly slower deterioration of the words

may be because the Keyword Mnemonic might benefit from the repeated retrieval of the caustic link and the imagery link when using Spaced Recall Practice. This strengthened imagery link as well as acoustic link in turn makes successful future retrieval possible via the Keyword Mnemonic, thus increasing the chances of successful L2 English word recall buy increasing the routs to such a recall.

CHAPTER FIVE

CONCLUSION

This study examined if there was a significant advantage to combining the Keyword Mnemonic method with Spaced Recall Practice as shown in the Combined Method. Forty-eight eighth grade students from a junior-high in Taichung city participated in the study. To control for the effect of using different classes, a within group experimental design was used. The students were sorted into three counterbalanced groups according to a stratified random sorting method. The stratified random sorting method involved dividing the students into a male and a female group and then ordering them into sets of three from highest to lowest ability levels based on their performances on the Vocabulary Levels Test before the experiment began. The individual students in each of these sets were then randomly shuffled into one of the three participant groups. Each participant group used both the Combined Method and Spaced Recall Practice to respectively study two lists of seven target English words from a three list set. The third list of words was not studied at all but was tested in the posttest. The list acted as the control. The three list set was rotated for each participant group so that each group had a different set of words for each of the three treatment conditions. The student participants were then again divided into three posttest groups using the stratified random sorting method again so

that an equal amount of participants from each participant group are in each of the posttest groups. Each posttest group then took the vocabulary retention test at one (and only one) of three time points (immediate posttest, first delayed posttest and second delayed posttest).

Repeated measures ANOVA analyses were performed to ascertain if there were any significant effects between the experimental conditions and the control condition for each posttest administered at three different time points Next, a one way ANOVA analysis was performed on all of the posttest results of the Combined Method condition to see if there were any significant effects between those results. Finally the same analysis was run to compare the posttest results of the Spaced Recall condition. This chapter will summarize the findings of the statistical analyses and discuss the pedagogical implications of the study. Finally, this chapter will conclude with the limitations of the study and suggestions for future research.

5.1 Summary of the major findings

The repeated measures ANOVA found no significant advantage for the Combined Method condition over the Spaced Recall condition in the results of the immediate posttest. A similar result was found in the first delayed posttest. No significant difference was found between the Combined Method condition and the Spaced Recall Practice condition using a repeated measures ANOVA and running a

post hoc t-tests. These results were in line with the findings of Fritz et al., (2007). However, the result of the repeated measures ANOVA did show a significant difference between the two experimental conditions and the control condition for both the immediate posttest and the first delayed posttest. These results suggest that the target L2 vocabulary was unknown to the participants of this study, eliminating prior knowledge as a confounding factor.

For the second delayed posttest, a repeated measures ANOVA showed that there was a significant difference between the three conditions. A post hoc analysis of the results confirmed that there were significant differences between all three conditions. Specifically the post hoc paired samples t-test did reveal a one tailed significant difference between the Combined Method condition and the Spaced Recall condition. This result implies that the Combined Method condition leads to the retention of significantly more words than the Spaced Recall Practice condition. This may also confirm that the rates of forgetting of the two treatment conditions are different. The fact that the initial posttests revealed no significant difference between the treatment conditions but one became apparent in the second delayed posttest confirms that these forgetting rates form curves when plotted on a graph.

A one way ANOVA was used to compare the means of the three posttests of the two experimental conditions separately. Significant differences were found among the

posttest mean scores for both the Combined Method condition and for the Spaced Recall condition. The results imply that for both conditions a significant amount of forgetting did take place between the immediate and the third (last) posttests. A post-hoc analysis of the results revealed that for the Combined Method condition there was only a significant difference between the immediate posttest and the third (last) posttest. In other words, the first delayed posttest results weren't significantly different from either the immediate or the third (last) posttest. For the Spaced Recall condition, the post-hoc analysis revealed a significant difference between the third (last) posttest and both the immediate and the first delayed posttest. There was, however, no significant difference between the first delayed posttest and the immediate posttest.

5.2 Pedagogical implications

As Dunlosky et al., (2013) pointed out, many effective learning techniques are underutilized while some less effective techniques are widely used. Thus, one of the goals of this study was to give educators additional information on effective vocabulary learning techniques. Specifically, it aimed to see if adding the Keyword Mnemonic to Spaced Recall Practice would be beneficial.

Although the Keyword Mnemonic has been shown to be effective in the retention of foreign vocabulary (Raugh, Schupbach, & Atkinson, 1977; Raugh &

Atkinson 1975; Atkinson & Raugh 1975; Atkinson, 1975), Wang, et al., (1993) showed that it was not effective for the long-term retention of vocabulary. However, Fritz, et al.,(2007) pointed out that the Keyword Mnemonic might be fragile when it came to leaving memory traces. In other words, even though it could be effective in the short term, these effects could be easily lost. They suggested that if the Keyword Mnemonic were combined with Spaced Recall Practice, and the treatment period were increased, this initial advantage may be retained.

In this study, it has been shown that there are benefits to combining the Keyword Method with Spaced Recall practice if certain criteria apply. For this combination to be effective, the words need to be spaced in learning sessions once a week over a period of at least a month. Furthermore, in this study the words were recalled three times in each learning session and immediate feedback was given. Thus, it is recommended that in the educational setting, the above mentioned criteria be used. This could be achieved using either projectors in the class and the students write the answers in their notebooks, or with the use of computers. However, educators would have to weigh the benefits up against the effort and time involved in creating the Keywords and imagery links for the Keyword Mnemonic.

The success of the Keyword Mnemonic depends a lot on not only the quality of keywords, but also the imagery links provided (Fritz et al., 2007). Thus, if keywords

and imagery links do not already exist, the educators need to not only come up with suitable keywords and imagery links, but also test their efficacy. This in itself may put many educators off from using the Keyword Mnemonic in combination with Spaced Recall Practice. Furthermore, not all words may be suitable for use with the keyword method (Dunlosky et al., 2013). In circumstances where it is either impossible to test the efficacy of the keywords and imagery links, or where the words are not suitable for use with the Keyword Mnemonic, it is recommended that Spaced Recall Practice be used instead of the Combined Method. This is because Spaced Recall Practice is a very effective technique in its own right (Dunlosky et al., 2013). Also, all the effort involved in creating suitable material for use with the Keyword Mnemonic may be better spent elsewhere if the certainty of its successful use is in doubt.

In conclusion, it is recommended to use the Combined Method in educational contexts for the memorization of vocabulary if the following prerequisites can be met. First, the target vocabulary should be retrieved from memory at least three times per learning session. Also, there should be intervals of around ten or more vocabulary words between each retrieval. Second, immediate feedback should be given. That is to say, after a word has been retrieved from memory, the correct target vocabulary should be provided for corrections to be made. Third, the ISI of the learning sessions should be around one week. Also, the learning sessions should take place over a

period of at least a month. Finally, if validated Keywords and imagery links are not available, the educators should create suitable Keywords and imagery links, and also effectively test their efficacy before using them in the classroom context.

5.3 Limitations of the study

Although the study did find a significant advantage to using the Combined Method over Spaced Recall Practice, there are some limitations to these findings.

Firstly, the participants in this study were EFL junior high students in Taiwan. Accordingly, the results may not generalize to other population groups such as English majors at the university level or ESL students in a country where English is regularly used in society outside of the educational context.

Secondly, the words in this study were selected on the basis of how concrete they were. Only nouns with concreteness ratings of 4.5 or more on a 5 point scale using the list developed by Brysbaert, et al. (2014) were used. The concreteness rating states how easily an image can be formed when thinking of a word (Raugh & Atkinson, 1975). The higher the concreteness rating for a word is the easier it is to imagine that word. Thus, words that are more abstract in nature may not perform as well, and if more of these words were on the lists used for this study, the results may have been different.

Thirdly, the current study only tested nouns. Laufer (1997b) stated that the part

of speech of a word may have an effect on how easy it is to learn the word, and that nouns seem to be the easiest to learn. Furthermore, Dunlosky et al., (2013) points out that certain words, such as nouns, may be more suitable for the Keyword Mnemonic, and, by implication, the Combined Method. Consequently, only using nouns in this study limits the generalizability of this study to nouns only.

Finally, only 42 participants were used in this study. Even though the groups were counter balanced using a stratified random sorting method to increase the statistical power, the small sample size needs to be taken into account when looking at the statistics and generalizing the results.

5.4 Suggestions for future research

Because of the limited number of participants in this study, the first suggestion for future research is to replicate the current study with a larger and more diverse sample size. This would strengthen the statistical power of the data analyses used in this study. Furthermore, such a study could also include a CALL condition instead of a control condition. Being able to demonstrate the usefulness of the Combined Method on a computer system would make it significantly easier to apply this method in the classroom setting, provided the computer resources exist. Also, although it seems like the Combined Method is significantly better for the memorization of

certain words when using paper and pencil, it cannot be assumed that this advantage will extend to CALL.

If the results of this study were replicated, the second suggestion for further research would be to compare the results of different parts of speech. As stated in the limitations of this study, only nouns were used. Thus, it is important to compare how well different parts of speech can be memorized using the Combined Method. This could yield important information regarding how well the Combined Method generalizes to parts of speech other than nouns.

Finally, as already mentioned, this study only tested nouns with concreteness ratings of 4.5 or more. Because this has a direct effect on the success of the Keyword Mnemonic, it is crucially important to investigate what this effect may be on the Combined Method This could be done in the following manner. First, words from the same part of speech need to be sorted into three groups with high (concreteness rating of 4.5 and higher), medium, (concreteness ratings between 3.25 and 2.75) and low (concreteness ratings lower than 0.5) concreteness. Then these three groups of words need to be memorized using the combined method and the results compared. The results could indicate what effect word concreteness has on the Combined Method.

Reference list

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, 11, 159–177.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Avila, E., & Sadoski, M. (1996). Exploring new applications of the keyword method to acquire English vocabulary. *Language Learning*, 46 (3), p. 379-395.

Atkinson, R. C. (1975). Mnemonics in second-language learning. *American Psychologist*, 30, 821-828.

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword to the acquisition of Russian vocabulary. *Journal of Experimental Psychology: Human Learning & Memory*, 104, 126-133.

Bacon, F., & In Devey, J. (1902). *Novum organum*. New York: P.F. Collier.

Baddeley, A. (1978). The trouble with levels: A re-examination of Craik and Lockhart framework for memory research. *Psychological Review*, 85, 139-152

Baddeley, A., & Longman, D. J. A. (1978). The influence of length and frequency on training sessions on the rate of learning to type. *Ergonomics*, 21, 627-635.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology*: General, 108, 296–308.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, 52, 566–577.

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over eight years. Journal of Experimental Psychology: *Learning, Memory and Cognition*, 13, 344–349.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? Cognitive Psychology, 61(3), 228–247. http://doi.org/10.1016/j.cogpsych.2010.05.004

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum

Bloom, K. F., & Shuell, T. J. (1981), Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, 74, 245-248.

Brown, T. S., & Perry, F. L., Jr. (1991). A comparison of three learning strategies for ESL vocabulary acquisition. *TESOL Quarterly*, 25, 655- 670.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40

thousand generally known English word lemmas. *Behavior Research Methods,* 46, 904-911.

Burman, D. D., Bitan, T., & Booth, J. R. (2008). Sex differences in neural processing of language among children. *Neuropsychologia*, 46(5), 1349–1362.

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

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 34, 918-928.

Campos, A., González, M. A., & Amor, A. (2003). Limitations of the mnemonic-keyword method. *The Journal of General Psychology*, 130(4), 399-413.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20(6), 633–642.

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

Clark, J.M., & Pavio, A. (2004). Extention of the Pavio, Yuille, and Madigan (1986) norms. *Behavior Research Methods,* 36(3), 371-383.

Dempster, F.N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79, 162-170.

Dempster, F. N. (1989). Spacing Effects and Their Implications for Theory and Practice. *Educational Psychology Review,* 1(4), 309-330.

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

Ebbinghaus, H. (1885) *Memory: A contribution to experimental psychology.* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover, 1964.

Folse, K. S. (2004). *The Underestimated Importance of Vocabulary in the Foreign Language Classroom*, [Online] Available: http://www.seasite.niu.edu/trans/articles/Underestimated%20Importance%20of%20Vocab.htm

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R. and Etkind, R. (2007), Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology,* 21: 499–526.

Gates, A. (1917). Recitation as a factor in memorizing. In R. Woodworth (Ed.), *Archives of Psychology* (Vol. 6, pp. 1-140). New York: Stanford University.

Gould, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*,11, 341-363.

Hall, J. W., Owens, W. L. & Wilson, K. P. (1987). Presentation rates and keywords in vocabulary learning. *Bulletin of the Psychonomic Society*, 25(3), 179-181.

Hill, M. & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Review of Applied Linguistics*, 41(2), 87-106.

Horst, M., Cobb, T. & Mera, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11(2) 207-223.

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.

Kang, S.H.K., Gollan, T.H., Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin and Review*, 20(6), 1259-1265.

Kang, S. H. K., Lindsey, R., Mozer, M. C., & Pashler, H. (2014). Retrieval practice

over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, 21, 1544–1550.

Karpicke, J. D. & Bauernschmidt, A. (2011). Spaced Recall : Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, 37(5), 1250-1257.

Karpicke J. D., Roediger H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally Spaced Recall enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704–719.

Karpicke J. D., Roediger H. L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151–162.

Karpicke J. D., Roediger H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, 38, 116–124.

Karpicke J. D., Blunt J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775.

Kraft, R. N., & Jenkins, J. J. (1981). The lag effect with aurally presented passages. *Bulletin of the Psychonomic Society*, 17, 132–134

Lancaster University (2009). DIALANG at Lancaster University. Retrieved from:

http://www.lancs.ac.uk/researchenterprise/dialang/about.htm.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). London: Academic Press.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, 43, 1174–1181.

Laufer, B. (1987). The lexical perspective of reading comprehension. *English Teachers' Journal (Israel)*, 35, 58-67.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint, & P. Arnaud, (Eds.), Vocabulary and Applied Linguistics, (pp.126-132). Macmillan.

Laufer, B. (1997a). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In M. McCarthy, & N. Schmitt, (eds.), *Vocabulary Description, Acquisition and Pedagogy*, (pp. 140-155). Cambridge University Press.

Laufer, B. (1997b). The lexical plight in second language reading: words you don't know, words you think you know and words you can't guess. In J. Coady, & T. Huckin, (Eds.), *Second Language Vocabulary Acquisition: a Rationale for*

Pedagogy, (pp. 20-34). Cambridge University Press.

Laufer, B. (2005). Focus on Form in second language vocabulary acquisition. In S. H. Foster-Cohen, M. P. Garcia-Mayo & J. C. Benjamins (Eds.), *EUROSLA Yearbook 5*, 223-250.

Laufer, B. & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning,* 48, 365–91.

Laufer, B. & Yano, Y. (2001). Understanding unfamiliar words in a text: do L2 learners understand how much they don't understand. *Reading in a Foreign Language* 13: 549-566.

Laufer, B. & Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. Language Learning 54, 469-523

Laufer, B. &Nation P. (1999) A Vocabulary Size Test of Controlled Productive Knowledge. *Language Testing,* 16(1), 33-51.

Lancaster University. (2015, July 22). Retrieved July 14, 2015, from http://www.lancaster.ac.uk/ researchenterprise/dialang/about

Logan, J. M. & Balota, D. A. (2008). Expanded vs. Equal Interval Spaced Recall Practice: Exploring Different Schedules of Spacing and Retention Interval in Younger and Older Adults. *Aging, Neuropsychology, and Cognition,* 15,

257-280.

Mastropieri, M. A. & Scruggs, T. E. (2012). Mnemonic Learning. In N. M. Steel, (Eds.), *Encyclopedia of the Sciences of Learning,* (pp.2287-2289). Springer US.

Meara. P. & Jones, G. (1989). *Eurocentres Vocabulary Test 10 KA.* Zurich: Eurocentres.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines* 5, 12–25.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary.* Boston, MA: Heinle and Heinle.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (2006a). How large a vocabulary is needed for reading and listening*? Canadian Modern Language Review*. 63 (1), 59-82.

Nation, I.S.P. (2006b) Language education - vocabulary. In K. Brown (ed.) *Encyclopaedia of Language and Linguistics*, 2nd Ed. vol6 (pp.494-499). Oxford: Elsevier.

Nation, I. S. P. & Waring, R. (1997), Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nation, I.S.P. & Beglar, D. (2007), A vocabulary size test. *The Language Teacher* 31 (7),

9-13.

Pashler, H., Cepeda, N. J.,Wixted, J. T.,& Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 3–8.

Pressley, M. (1987). Are keyword method effects limited to slow presentation rates? An empirically-based reply to Hall and Fuson (1986). *Journal of Educational Psychology* , 79, 333-335.

Pressley, M. (1988). A review of studies establishing that keyword method effects are robust with respect to temporal characteristics of presentation, *Bulletin of the Psychonomic Society,* 26(2), 94-96.

Pressley, M., & Levin, J. R. (1978). Development constraints associated with children's use of the keyword method of foreign language vocabulary learning. *Journal of Experimental Child Psychology*, 26, 359-372.

Pyc, M.A. & Rawson, K.A. (2009) Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Launguage*, 60, 437–447.

Raugh, M. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second language vocabulary. *Journal of Educational Psychology*, 67, 1–16.

Raugh, M. R., Schupbach, R. D., & Atkinson, R. C. (1977). Teaching a large Russian

language vocabulary by the mnemonic keyword method. *Instructional Science*, 6, 199–221.

Rodriguez, M. & Sadoski, M. (2000). Effects of Rote, Context, Keyword, and Context/keyword Methods on Retention of Vocabulary in EFL Classrooms. *Language Learning*. 50 (2), 385-412.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15, 20–27.

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

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual.* Basingstoke, England: Palgrave Macmillan.

Schmitt, N. Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal,* 95(1), 26–43.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, 30, 641–656.

Swanborn, M. S. L. & de Glopper, K. (2002), Impact of Reading Purpose on Incidental Word Learning from Context. *Language Learning*, 52: 95–117

Wang, A., Thomas, M., Inzana, C., & Primicerio, L. (1993). Long-term retention under conditions of intentional learning and the keyword mnemonic. *Bulletin of the Psychonomic Society*. 31(6), 545–54

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, 1–27.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review*, 53, 13-40.

Yates, F. A., (1966). *The art of memory*. Chicago: University of Chicago Press

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, 24, 39-58.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, 57(4), 541-572.

Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A., & Healy, N.A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27(2),

201–212.

Appendix A

List of English-Chinese word Pairs used in study

L2 Target words	L1 Equivalents	Keywords	Descriptive phrases	#
spike	牆頭釘，尖鐵	死台客	一個被釘子穿刺過的死台客	1
creek	小河，溪	貴客	有一個貴客站在小溪邊	2
lava	熔岩	蛤蜊 (台語)	在熔岩中燃燒的蛤蜊	3
comet	彗星	摳門	他太摳門了，所以不願意花錢買彗星的照片	4
sleigh	雪橇	死累	一邊快累死了地一邊划雪橇	5
seasoning	調味料	犧牲你	我會為了一些調味料而犧牲你	6
mule	騾	木耳	有一隻騾身上長了木耳	7
shield	盾	洗耳朵	常常清洗你的耳朵就像對感染持有擋箭牌一樣	8
crust	麵包皮，派餅皮	快死了	在吃麵包皮時嗆到快死了	9
oar	槳，櫓	偶爾	偶爾用槳划船	10
blade	刀片	布類的	這種刀片適合裁剪布類的東西	11
jockey	賽馬的騎師	很生氣 (台語)	那個賽馬的騎師非常生氣的鞭打他的馬	12
daffodil	水仙花	大糞多	小心那個種水仙花田大糞多	13
pillar	柱，柱子	皮了	頑皮的小男孩把柱子推倒了	14
vein	靜脈，血管	麵	有人在吃血管像吃麵條一樣	15
eel	鰻魚	咿唷	驢子咿唷騎在一條鰻魚上	16
hood	帽兜	虎的	那件衣服的帽兜上有虎的圖案	17
hoof	蹄	護膚	吃豬蹄可以護膚	18
pickle	l醃漬食品；醃菜	劈狗	他在劈一條狗來做成醃肉	19
celery	芹菜	殺了你	你不把那些芹菜吃光我就會殺了你	20
mermaid	美人魚	萌妹	那個萌妹常幻想變成一條美人魚	21

Appendix B
Traditional Chinese version of Nation's Vocabulary Levels Test

Second 1000
1.
maintain: Can they maintain it?
a.
維持
b.
擴大
c.
改善
d.
得到
2.
stone: He sat on a stone.
a.
石頭
b.
凳子
c.
墊子
d.
樹枝
3.
upset: I am upset.
a.
疲倦的
b.
著名的
c.
富足的
d.
不高興的
4.
drawer: The drawer was empty.
a.
抽屜
b.
車庫
c.
冰箱
d.
鳥籠
5.
patience: He has no patience.
a.
沒有耐心
b.
很忙
c.
沒有信心
d.
不公正
6.
nil: His mark for that question was nil.
a.
很差的
b.
什麼也沒有的
c.
很好的
d.
中等的
7.
pub: They went to the pub.
a.
酒吧
b.
銀行
c.
商場
d.
游泳池
8.
circle: Make a circle.
a.
素描
b.
空白
c.
圓圈
d.
大洞
9.
microphone: Please use the microphone.
a.
微波爐
b.
麥克風
c.
顯微鏡
d.
手機
10.
pro: He's a pro.
a.
間諜
b.
傻瓜
c.
記者
d.
職業運動員

Second 1000
1.
maintain: Can they maintain it?
a.
維持
b.
擴大
c.
改善
d.
得到
2.
stone: He sat on a stone.
a.
石頭
b.
凳子
c.
墊子
d.
樹枝
3.
upset: I am upset.
a.
疲倦的
b.
著名的
c.
富足的
d.
不高興的
4.
drawer: The drawer was empty.
a.
抽屜
b.
車庫
c.
冰箱
d.
鳥籠
5.
patience: He has no patience.
a.
沒有耐心
b.
很忙
c.
沒有信心
d.
不公正
6.
nil: His mark for that question was nil.
a.
很差的
b.
什麼也沒有的
c.
很好的
d.
中等的
7.
pub: They went to the pub.
a.
酒吧
b.
銀行
c.
商場
d.
游泳池
8.
circle: Make a circle.
a.
素描
b.
空白
c.
圓圈
d.
大洞
9.
microphone: Please use the microphone.
a.
微波爐
b.
麥克風
c.
顯微鏡
d.
手機
10.
pro: He's a pro.
a.
間諜
b.
傻瓜
c.
記者
d.
職業運動員

Third 1000
1.
soldier: He is a soldier.
a.
商人
b.
學生
c.
金屬工藝制造者
d.
士兵
2.
restore: It has been restored.
a.
重復
b.
重新分配
c.
降價
d.
復原
3.
jug: He was holding a jug.
a.
罐子
b.
聊天
c.
貝雷帽
d.
槍
4.
scrub: He is scrubbing it.
a.
抓
b.
修理
c.
刷洗
d.
作素描
5.
dinosaur: The children were pretending to be dinosaurs.
a.
海盜
b.
仙女
c.
龍
d.
恐龍
6.
strap: He broke the strap.
a.
諾言
b.
蓋子
c.
盤子
d.
帶子
7.
pave: It was paved.
a.
堵塞
b.
分開
c.
鑲金邊
d.
鋪路
8.
dash: They dashed over it.
a.
猛沖
b.
磨蹭
c.
爭吵
d.
瞥見
9.
rove: He couldn't stop roving.
a.
喝醉
b.
漂泊
c.
哼曲子
d.
努力工作
10
lonesome: He felt lonesome.
a.
不領情的
b.
疲倦的
c.
孤獨的
d.
精力充沛的

Fourth 1000
1.
compound: They made a new compound.
a.
協議
b.
復合物
c.
公司
d.
預言
2.
latter: I agree with the latter.
a.
牧師
b.
理由
c.
后者
d.
答案
3.
candid: Please be candid.
a.
小心的
b.
表示同情的
c.
公平的
d.
直率的
4.
tummy: Look at my tummy.
a.
圍巾
b.
肚子
c.
鬆鼠
d.
拇指
5.
quiz: We made a quiz.
a.
箭筒
b.
錯誤
c.
競賽
d.
鳥巢
6.
input: We need more input.
a.
輸入
b.
工人
c.
填料
d.
錢
7.
crab: Do you like crabs?
a.
蟹
b.
薄脆餅干
c.
又緊又硬的領子
d.
蟋蟀
8.
vocabulary: You will need more vocabulary.
a.
詞匯
b.
技巧
c.
錢
d.
槍
9.
remedy: We found a good remedy.
a.
矯正問題的方法
b.
餐館
c.
食譜
d.
等式
10.
allege: They alleged it.
a.
辯解
b.
剽竊
c.
証明
d.
反抗

Fifth 1000
1.
deficit: The company had a large deficit.
a.
出現赤子
b.
貶值
c.
有這筆大開銷的計劃
d.
在銀行裡有很多存款
2.
weep: He wept.
a.
畢業
b.
哭
c.
死
d.
擔心
3.
nun: We saw a nun.
a.
蠕虫
b.
事故
c.
修女
d.
天空中無法解釋的亮光
4.
haunt: The house is haunted.
a.
充滿了裝飾物
b.
已被出租
c.
空的
d.
鬧鬼
5.
compost: We need some compost.
a.
大力支持
b.
扶持
c.
混凝土
d.
堆肥
6.
cube: I need one more cube.
a.
大頭針
b.
立方體
c.
缸子
d.
卡片
7.
miniature: It is a miniature.
a.
微型畫
b.
顯微鏡
c.
微生物
d.
在書法中把字母連在一起細小的連線
8.
peel: Shall I peel it?
a.
浸泡
b.
削皮
c.
燙洗
d.
切成薄片
9.
fracture: They found a fracture.
a.
裂口
b.
碎片
c.
夾克衫
d.
稀有的寶石
10.
bacterium: They didn't find a single bacterium.
a.
細菌
b.
開有紅色或橘黃色花的植物
c.
駱駝
d.
贓物

Sixth 1000
1.
devious: Your plans are devious.
a.
詭計多端的
b.
成熟的
c.
考慮不周詳的
d.
過於昂貴的
2.
premier: The premier spoke for an hour.
a.
律師
b.
講師
c.
冒險家
d.
總理
3.
butler: They have a butler.
a.
男管家
b.
鋸
c.
家庭教師
d.
地窖
4.
accessory: They gave us some accessories.
a.
簽証
b.
官方命令
c.
選擇
d.
額外的零件
5.
threshold: They raised the threshold.
a.
旗子
b.
門檻
c.
天花板
d.
利息
6.
thesis: She has completed her thesis.
a.
論文
b.
歸納証詞
c.
試用期
d.
延期治療
7.
strangle: He strangled her.
a.
掐死
b.
寵壞
c.
綁架
d.
贊美
8.
cavalier: He treated her in a cavalier manner.
a.
慢待的
b.
禮貌的
c.
尷尬的
d.
兄長的
9.
malign: His malign influence is still felt.
a.
邪惡的
b.
好的
c.
非常重要的
d.
秘密的
10.
veer: The car veered.
a.
改變方向或路線
b.
劇烈晃動
c.
發生逆火引起爆鳴
d.
打滑

Seventh 1000
1.
olive: We bought olives.
a.
橄欖
b.
康乃馨
c.
男人的游泳衣
d.
清除雜草的工具
2.
quilt: They made a quilt.
a.
遺囑
b.
合同
c.
被子
d.
羽毛筆
3.
stealth: They did it by stealth.
a.
花費大量的錢
b.
逼迫
c.
悄悄的或秘密的行動
d.
沒有注意到所遇到的問題
4.
shudder: The boy shuddered.
a.
低語
b.
差點摔倒
c.
發抖
d.
大聲叫喊
5.
bristle: The bristles are too hard.
a.
問題
b.
短而硬的毛發
c.
折疊床
d.
鞋底
6.
bloc: They have joined this bloc.
a.
樂隊
b.
小偷幫
c.
偵察員
d.
集團
7.
demography: This book is about demography.
a.
土地使用模式研究
b.
用圖片表示數字事實的研究
c.
水文學
d.
人口學
8.
gimmick: That's a good gimmick.
a.
高空作業時所站的東西
b.
錢包
c.
引人注意的行為或事物
d.
花招
9.
azalea: This azalea is very pretty.
a.
杜鵑花
b.
由天然棉所制成的很輕的材料
c.
莎麗
d.
扇貝
10.
yoghurt: This yoghurt is disgusting.
a.
淤泥
b.
傷口
c.
酸奶
d.
榲桲

Eighth 1000
1.
erratic: He was erratic.
a.
完美的
b.
很壞的
c.
很有禮貌的
d.
不可靠的
2.
palette: He lost his palette.
a.
裝魚的籃子
b.
胃口
c.
年輕的女伴
d.
調色板
3.
null: His influence was null.
a.
具有好的結果
b.
毫無幫助的
c.
沒有效果的
d.
持久的
4.
kindergarten: This is a good kindergarten.
a.
消遣
b.
幼兒園
c.
背包
d.
圖書館
5.
eclipse: There was an eclipse.
a.
颶風
b.
潑濺
c.
大屠殺
d.
日食
6.
marrow: This is the marrow.
a.
吉祥物
b.
骨髓
c.
操縱杆
d.
增加工資
7.
locust: There were hundreds of locusts.
a.
飛蝗
b.
志願者
c.
素食者
d.
顏色鮮艷的野花
8.
authentic: It is authentic.
a.
真的
b.
非常吵鬧的
c.
老的
d.
干旱的
9.
cabaret: We saw the cabaret.
a.
壁畫
b.
卡巴萊
c.
蟑螂
d.
美人魚
10.
mumble: He started to mumble.
a.
集中精力
b.
顫抖
c.
遠遠地落后於其他人
d.
咕噥

Ninth 1000
1.
hallmark: Does it have a hallmark?
a.
表明什麼時候之前應該被使用的印記
b.
純度印記
c.
表明經過皇室同意的標記
d.
禁止復制的標記
2.
puritan: He is a puritan.
a.
喜歡被人注意的人
b.
具有嚴格道德標准的人
c.
吉普賽人
d.
守財奴
3.
monologue: Now he has a monologue.
a.
單眼鏡
b.
獨白
c.
專制
d.
把字母有趣地連在一起的圖畫
4.
weir: We looked at the weir.
a.
行為古怪的人
b.
紅樹屬植物
c.
通過吹來演奏的古老金屬樂器
d.
攔河壩
5.
whim: He had lots of whims.
a.
古老的金幣
b.
母馬
c.
沒有任何動機的奇異想法
d.
疼痛的紅腫塊
6.
perturb: I was perturbed.
a.
被迫同意的
b.
煩惱的
c.
困惑的
d.
濕透的
7.
regent: They chose a regent.
a.
不負責人的人
b.
暫時主持會議的人
c.
攝政者
d.
代表
8.
octopus: They saw an octopus.
a.
貓頭鷹
b.
潛水艇
c.
直升飛機
d.
章魚
9.
fen: The story is set in the fens.
a.
沼澤
b.
山地
c.
貧民窟
d.
很久以前
10.
lintel: He painted the lintel.
a.
過梁
b.
渡船
c.
長有伸展樹枝和綠色果實的美麗的樹
d.
戲院中顯示場景的板子

Tenth 1000
1.
awe: They looked at the mountain with awe.
a.
擔心
b.
興趣
c.
驚奇
d.
尊重
2.
peasantry: He did a lot for the peasantry.
a.
當地人
b.
寺廟
c.
商人俱樂部
d.
農民
3.
egalitarian: This organization is very egalitarian.
a.
保密的
b.
保守的
c.
訴訟的
d.
平等主義的
4.
mystique: He has lost his mystique.
a.
體格
b.
神秘性
c.
情人
d.
胡子
5.
upbeat: I'm feeling really upbeat about it.
a.
苦惱的
b.
樂觀的
c.
受傷害的
d.
迷惑的
6.
cranny: We found it in the cranny!
a.
舊雜物義賣
b.
小洞
c.
閣樓
d.
大箱子
7.
pigtail: Does she have a pigtail?
a.
辮子
b.
長的衣、袍、裙等拖在地上的部分
c.
開有一串下垂淺粉色花的植物
d.
情人
8.
crowbar: He used a crowbar.
a.
撬棍
b.
化名
c.
錐子
d.
輕的金屬拐杖
9.
ruck: He got hurt in the ruck.
a.
骨盆
b.
打架
c.
（運動員或競賽者等的）散亂一群
d.
在雪地上奔跑
10.
lectern: He stood at the lectern.
a.
講台
b.
聖餐桌
c.
酒吧
d.
邊緣

Eleventh 1000
1.
excrete: This was excreted recently.
a.
排泄
b.
澄清
c.
調查
d.
被列入不合法的事情之中
2.
mussel: They bought mussels.
a.
玻璃彈球
b.
貽貝
c.
榲桲
d.
紙巾
3.
yoga: She has started yoga.
a.
粗繩結的飾物
b.
瑜伽
c.
羽毛球
d.
東方國家的一種舞蹈
4.
counterclaim: They made a counterclaim.
a.
在法律案件中一方所提出的訴求和另一方的訴求相同
b.
要求商店收回有瑕疵的東西
c.
兩個公司之間交換工作的合同
d.
被子
5.
puma: They saw a puma.
a.
由坯建成的小房子
b.
來自炎熱而干旱國家的樹
c.
颶風
d.
美洲獅
6.
pallor: His pallor caused them concern.
a.
非正常的高溫
b.
對一切都毫無興趣
c.
一幫朋友
d.
蒼白的皮膚
7.
aperitif: She had an aperitif.
a.
躺椅
b.
家庭教師
c.
帶有很高羽毛的大帽子
d.
開胃酒
8.
hutch: Please clean the hutch.
a.
隔板
b.
行李箱
c.
輪轂
d.
兔籠
9.
emir: We saw the emir.
a.
尾巴上長有兩個長長卷曲羽毛的鳥
b.
奶媽
c.
埃米爾（對穆斯林統治者的尊稱）
d.
（愛斯基摩人用堅硬雪快砌成的臨時棲身用的）拱型圓頂小屋
10.
hessian: She bought some hessian.
a.
光澤油亮略呈粉紅色的魚
b.
大麻
c.
一種結實的粗麻布
d.
用來給食物調味的味道濃烈的根狀物

Twelfth 1000
1.
haze: We looked through the haze.
a.
舷窗
b.
薄霧
c.
窗帘
d.
花名冊
2.
spleen: His spleen was damaged.
a.
膝蓋骨
b.
脾臟
c.
下水管
d.
自尊
3.
soliloquy: That was an excellent soliloquy!
a.
六個人唱的歌曲
b.
修飾詞或描述性詞語
c.
帶有燈光和音樂的娛樂
d.
獨白
4.
reptile: She looked at the reptile.
a.
手稿
b.
爬行動物
c.
挨家兜售貨物的人
d.
水粉畫
5.
alum: This contains alum.
a.
取自一種普通植物的有毒物質
b.
一種由人造棉所制成的柔軟材料
c.
鼻煙
d.
白礬
6.
refectory: We met in the refectory.
a.
食堂
b.
簽署法律文件的辦公室
c.
宿舍
d.
溫室
7.
caffeine: This contains a lot of caffeine.
a.
一種讓人瞌睡的物質
b.
由堅韌樹葉織成的線狀物
c.
錯誤的觀點
d.
咖啡因
8.
impale: He nearly got impaled.
a.
控告
b.
監禁
c.
被尖物刺中
d.
陷入爭論之中
9.
coven: She is the leader of a coven.
a.
合唱隊
b.
集體企業
c.
秘密組織
d.
過著嚴格宗教生活的一群生活在教堂的女人
10.
trill: He practised the trill.
a.
顫音
b.
小提琴
c.
投球
d.
(跳芭蕾舞者的)單足旋轉

Thirteenth 1000
1.
ubiquitous: Many weeds are ubiquitous.
a.
很難鏟除的
b.
長有又長又壯根的
c.
在大部分國家都能夠被發現的
d.
在冬天枯竭的
2.
talon: Just look at those talons!
a.
山頂
b.
爪
c.
盔甲
d.
傻瓜
3.
rouble: He had a lot of roubles.
a.
紅寶石
b.
親戚
c.
盧布
d.
心中的道德或者其他難題
4.
jovial: He was very jovial.
a.
社會地位低下的
b.
總愛批評人的
c.
很幽默的
d.
很友好的
5.
communiqué: I saw their communiqué.
a.
對一個機構的批評報告
b.
屬於社區成員的花園
c.
用於做廣告的印刷材料
d.
官方通告
6.
plankton: We saw a lot of plankton.
a.
毒性傳播很快的雜草
b.
生活在水中的很小的植物或動物
c.
堅硬木質的樹木
d.
容易導致土地滑坡的灰色泥土
7.
skylark: We watched a skylark.
a.
飛行表演
b.
人造衛星
c.
表演魔術的人
d.
一邊鳴叫一邊飛得很高的小鳥
8.
beagle: He owns two beagles.
a.
車頂能夠收起來的跑得很快的車
b.
能夠快速射中很多人的槍
c.
長有長長耳朵的小狗
d.
建在度假勝地的房子
9.
atoll: The atoll was beautiful.
a.
環狀珊瑚島
b.
用細線織成圖畫的藝術品
c.
女人晚上所佩帶的鑲嵌有寶石的小皇冠
d.
河水流過大石間的縫隙
10.
didactic: The story is very didactic.
a.
說教的
b.
難以置信的
c.
令人興奮的
d.
令讀者琢磨不透的

Fourteenth 1000
1.
canonical: These are canonical examples.
a.
打破常規的
b.
取自一本宗教書的
c.
合乎原則並被廣泛接受的
d.
最近發現的
2.
atop: He was atop the hill.
a.
在… 腳下
b.
在… 頂部
c.
在這邊
d.
在那邊
3.
marsupial: It is a marsupial.
a.
長有堅硬腳的動物
b.
生長幾年的一種植物
c.
開出的花總是面向太陽的植物
d.
有袋類動物
4.
augur: It augured well.
a.
預示未來的好事情
b.
和預料的很吻合
c.
具有一種和其它東西搭配在一起很漂亮的色彩
d.
發出一種清脆而動聽的聲音
5.
bawdy: It was very bawdy.
a.
難以預料的
b.
令人愉快的
c.
急促的
d.
粗魯的
6.
gauche: He was gauche.
a.
善談的
b.
靈活的
c.
尷尬的
d.
堅決的
7.
thesaurus: She used a thesaurus.
a.
一種字典
b.
一種化合物
c.
一種特殊的說話方式
d.
皮下注射
8.
erythrocyte: It is an erythrocyte.
a.
止痛的藥
b.
血液中的紅色的成分
c.
略顯紅色的白金屬
d.
鯨家族中的一員
9.
cordillera: They were stopped by the cordillera.
a.
特殊的法律
b.
裝備有武器的船隻
c.
山脈
d.
國王的長子
10.
limpid: He looked into her limpid eyes.
a.
清澈的
b.
含淚的
c.
深棕色的
d.
美麗的

Appendix C

Answer Key and Test Paper of Vocabulary Retention Test

Please fill in the English equivalent words in the column next to the Chinese words.

Name_____Student Number (class plus number)_____

1. 牆頭釘，尖鐵	1.spike
2. 小河，溪	2.creek
3. 熔岩	3.lava
4. 彗星	4.comet
5. 雪橇	5.sleigh
6. 調味料	6.seasoning
7. 騾	7.mule
8. 盾	8.shield
9. 麵包皮，派餅皮	9.crust
10. 槳，櫓	10.oar
11. 刀片	11.blade
12. 賽馬的騎師	12.jockey
13. 水仙花	13 .daffodil
14. 柱，柱子	14. pillar
15. 靜脈，血管	15. vein
16. 鰻魚	16. eel
17. 帽兜	17. hood
18. 蹄	18. hoof
19. 1醃漬食品；醃菜	19. pickle
20. 芹菜	20. celery
21. 美人魚	21. mermaid

Vocabulary Retention Test: Answer Key

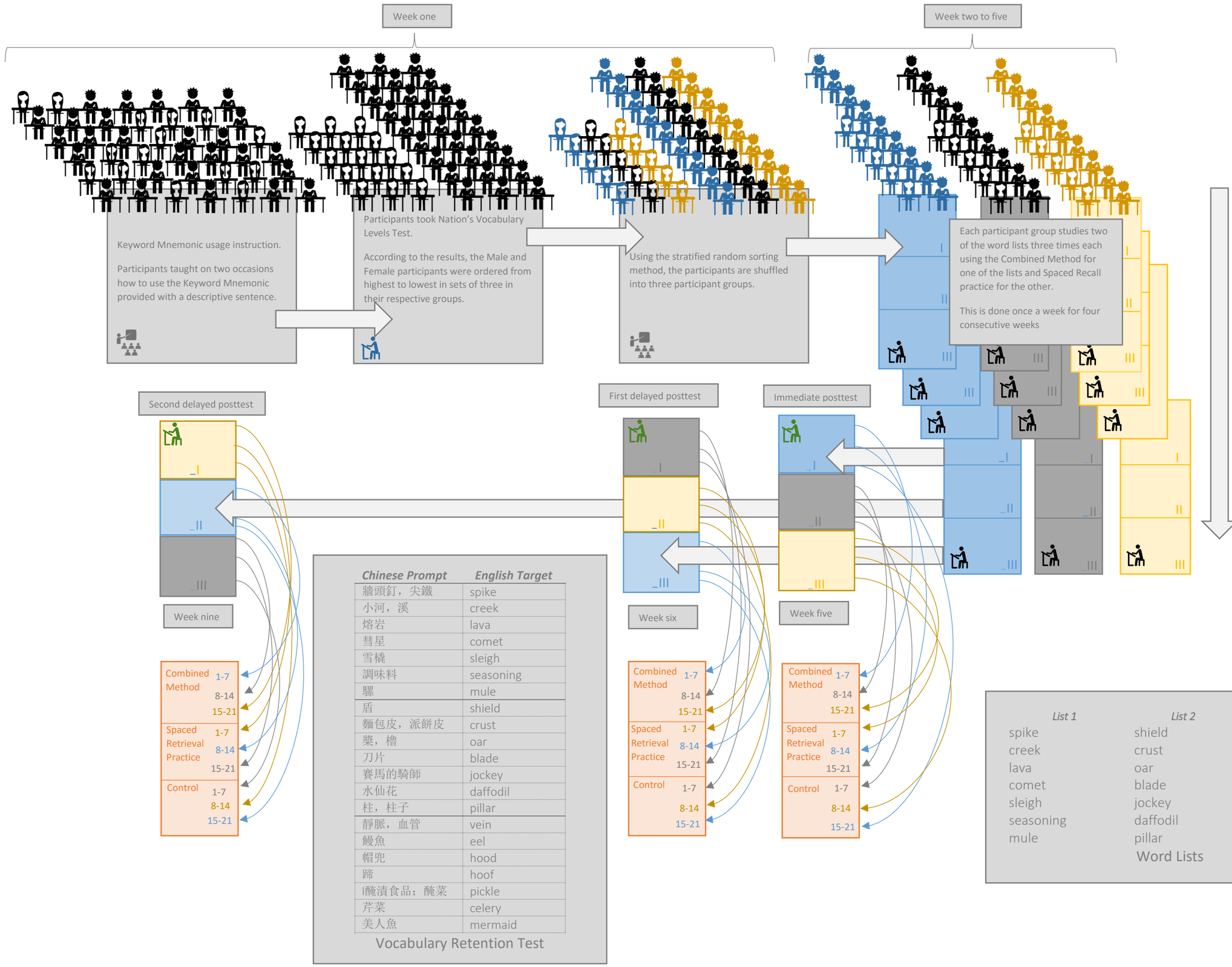
Please fill in the English equivalent words in the column next to the Chinese words.

Name_____Student Number (class plus number)_____

1. 牆頭釘，尖鐵	1.
2. 小河，溪	2.
3. 熔岩	3.
4. 彗星	4.
5. 雪橇	5.
6. 調味料	6.
7. 騾	7.
8. 盾	8.
9. 麵包皮，派餅皮	9.
10. 槳，櫓	10.
11. 刀片	11.
12. 賽馬的騎師	12.
13. 水仙花	13.
14. 柱，柱子	14.
15. 靜脈，血管	15.
16. 鰻魚	16.
17. 帽兜	17.
18. 蹄	18.
19. 1醃漬食品；醃菜	19.
20. 芹菜	20.
21. 美人魚	21.

Vocabulary Retention Test: Test Paper

APPENDX D


Week one
Week two to five
Keyword Mnemonic usage instruction.
Participants taught on two occasions how to use the Keyword Mnemonic provided with a descriptive sentence.
Participants took Nation's Vocabulary Levels Test.
According to the results, the Male and Female participants were ordered from highest to lowest in sets of three in their respective groups.
Using the stratified random sorting method, the participants are shuffled into three participant groups.
Each participant group studies two of the word lists three times each using the Combined Method for one of the lists and Spaced Recall practice for the other.
This is done once a week for four consecutive weeks
Second delayed posttest
First delayed posttest
Immediate posttest
Week nine
Week six
Week five
Combined Method 1-7 8-14 15-21
Spaced Retrieval Practice 1-7 8-14 15-21
Control 1-7 8-14 15-21
Combined Method Spaced Retrieval Practice Control
spike shield
creek crust
lava oar
comet blade
sleigh jockey
seasoning daffodil
mule pillar
Word Group A: List 1 CM List 2 SRP
shield vein
crust eel
oar hood
blade hoof
jockey pickle
daffodil celery
pillar mermaid
Word Group B: List 2 CM List 3 SRP
vein shield
eel crust
hood oar
hoof blade
pickle jockey
celery daffodil
mermaid pillar
Word Group C: List 3 CM List 1 SRP
Chinese Prompt | English Target
牆頭釘，尖鐵 | spike
小河，溪 | creek
熔岩 | lava
彗星 | comet
雪橇 | sleigh
調味料 | seasoning
騾 | mule
盾 | shield
麵包皮，派餅皮 | crust
槳，櫓 | oar
刀片 | blade
賽馬的騎師 | jockey
水仙花 | daffodil
柱，柱子 | pillar
靜脈，血管 | vein
鰻魚 | eel
帽兜 | hood
蹄 | hoof
l醃漬食品；醃菜 | pickle
芹菜 | celery
美人魚 | mermaid
Vocabulary Retention Test
List 1 List 2 List 3
spike shield vein
creek crust eel
lava oar hood
comet blade hoof
sleigh jockey pickle
seasoning daffodil celery
mule pillar mermaid
Word Lists