Investigating the effect of spaced versus massed practice on vocabulary retention in the EFL classroom

Ahmad Al Fotais

A thesis submitted for the degree of

Doctorate of Philosophy

in

Applied Linguistics

Department of Language and Linguistics

University of Essex

June 2019

Abstract

The expression 'spacing effect' refers to a commonly observed finding that spacing learning over a period of time leads to better retention than massing learning in a single session. The present study for the first time experimentally compared the relative effectiveness of spaced practice and massed practice on vocabulary learning in authentic EFL classroom settings at tertiary level. This thesis examined the difference in initial learning and longer-term retention between massed and spaced practice at four strength levels of knowledge of vocabulary meaning, namely receptive recognition (easiest), productive recognition, receptive recall, and productive recall (hardest) (Laufer & Goldstein, 2004). Furthermore, this thesis examined the difference in initial learning and retention between word classes, the role of individual factors in spaced learning compared with massed learning, and whether the four levels of vocabulary strength additionally constituted an implicational scale.

With these aims, year-one Saudi EFL university students were taught the meaning of 30 new words in a massed learning condition and 30 other new words in a spaced learning condition. In the massed learning condition, each target word was practiced four times in one classroom session. In the spaced learning condition, each target word was practiced once in each of four classroom sessions. The same vocabulary tests were administered immediately after the intervention and four weeks later. Questionnaires were additionally used to gather self-reported individual data.

The findings revealed that scores for items that were learnt in the massed condition were not only lower than scores for items that were learnt in the spaced condition but also yielded a greater fall between the immediate and delayed post-tests, although that fall was not significant at the easiest strength level. The benefit of spaced learning over massed learning applied equally to nouns and verbs, with the former's scores being higher regardless of the time when the test occurred. Vocabulary learning with spaced practice was beneficial to all learners irrespectively of whether they preferred it or not over massed practice. The study agrees with Laufer and Goldstein's (2004) finding of an implicational scale across the same four degrees of knowledge strength. In addition to further results, implications for second language acquisition and vocabulary learning theory, and English as a foreign language pedagogy are presented.

Acknowledgments

I thank Allah the Almighty for providing me with guidance, patience, time, and strength to have this work carried out. Prophet Mohammad (peace be upon him) said “He will not be thankful to Allah, he who would not be thankful to people” (correctly reported by Tirmidhi). I would like to dedicate the following paragraphs to those who helped me in many ways along the way of making this thesis.

First, my gratitude goes to my supervisor Dr. Sophia Skoufaki for nurturing my early interest in the topic and for providing me with her invaluable feedback, thoughts, and expertise in the field of vocabulary research. I am also indebted to Mr. Phillip Scholfield for his assistance and encouragement during the critical time of my study. My profound thanks are also due to Professor Monika Schmid, Professor Florence Myles, Dr. Karen Roehr-Brackin, Dr. Christina Gkonou, Dr. Kamal Alsharfi, Dr. Sami Althumali, and Dr. Ahmed Al-Masrai for their contributions to this thesis and the great influence they have had on my professional training.

My wholehearted gratitude goes to my parents for their care and support throughout my life. They nurtured in me the passion for learning and without them, I would not be the man I am today. Thank you with all my heart and forgive me for being away from you all those years.

Words are too humble to express my gratitude to my beloved wife Asma AlQahtani. Her patience, understanding, sacrifice, support, and encouragements made this thesis possible. She will always be a part in any success I achieve in life. I am also blessed with my beautiful children Reem and Mishary as their smiles made all difficulties fade away and made me work harder and persevere in my studies.

Finally, I am grateful to Taif University which granted me a scholarship to undertake my PhD. I also acknowledge the vital contribution of the research participants, without whom none of this research could have taken place. I am also grateful to all family and friends for their moral support and prayers. I like to also thank all members of staff at the department of language and linguistics in Essex University for their feedback and valuable talks in seminars and conferences organized by the department.

List of Abbreviations

BA	Bachelor of Art
CATSS	Computer Adaptive Test of Size and Strength
EFL	English as a Foreign Language
ELT	English Language Teaching
FL	Foreign Language
H	Hypothesis
KSA	Kingdom of Saudi Arabia
L1	First Language
L2	Second Language
POS	Part of Speech
PVLT	Productive Vocabulary Level Test
RQ	Research Question
SLA	Second Language Acquisition
TU	Taif University
VATS	Vocabulary Achievement Test of Strength
VLT	Vocabulary Level Test

Table of Contents

Abstract ii

Acknowledgments iii

List of Abbreviations iv

List of Tables x

List of Figures xii

Chapter One - Introduction 1

1.1 Research aim 1

1.2 Background to the study 1

1.2.1 The problem of EFL vocabulary in the Saudi context 1

1.2.2 Gaps in the study of repetition/recycling in foreign language vocabulary learning and teaching in general 7

1.3 Methodology 11

1.4 Delimitation of the research 12

1.5 Thesis outline 12

Chapter Two - Literature Review 14

2.1 Introduction 14

2.2 The importance of vocabulary learning in foreign language learning 15

2.3 What does *word* mean? 18

2.3.1 Tokens and types 20

2.3.2 Lemmas and word families 22

2.4 The nature of vocabulary knowledge 27

2.5 Instructed second and foreign language vocabulary learning 31

2.6 Factors potentially affecting vocabulary learnability that are relevant to the present study 39

2.6.1 Word class 40

2.6.2 Word length 41

2.6.3 L1 lexical translation equivalence 43

2.6.4 Frequency of occurrence 49

2.6.5 Recycling and Repetition 53

2.7 The spacing effect 55

2.7.1 Theoretical accounts for the spacing effect 56

2.8 Research on the spacing effect on vocabulary retention 57

2.9 Classroom-based research into the spacing effect on vocabulary retention 61

2.9.1 Sobel et al. (2011)..62

2.9.2 Goossens et al. (2012) 63

2.9.3 Schuetze (2015)...65

2.9.4 Limitations of previous studies 66

2.10 Vocabulary testing 68

2.10.1 Multiple measures of vocabulary knowledge (depth)...................................70

2.10.2 Vocabulary size tests (breadth) 73

2.10.3 Yes/No tests...75

2.10.3.1 The orthographic vocabulary size test (X-Lex)...77

2.10.3.2 The aural vocabulary size test (A_lex)...77

2.10.3.3 Eurocentres Vocabulary Size Test (EVST)...78

2.10.4 Multiple-choice tests..79

2.10.4.1 Vocabulary Levels Test (VLT)...80

2.10.4.2 Vocabulary Size Test (VST)...82

2.10.5 Translation tests 84

2.10.5.1 Computer Adaptive Test of Size and Strength (CATSS) 85

2.11 Conclusion 90

2.12 Research questions and hypotheses concerning effects of spacing on vocabulary learning 92

Chapter Three - Methodology 94

3.1 Introduction 94

3.2 Objectives 95

3.3 Research paradigm..96

3.4 Main study participants 98

3.5 Design..99

3.6 Materials for the main study massed/spaced intervention sessions 103

3.6.1 Vocabulary selection 103

3.6.2 Teaching phase activities...107

3.7 Instrumentation 109

3.7.1 Questionnaires...110

3.7.1.1 Biodata and language learning history questionnaire..110

3.7.1.2 Learners’ perception questionnaire...111

3.7.2 Vocabulary Size Test (VST) (Nation & Beglar, 2007)......................................112

3.7.3 Design of Vocabulary Achievement Test of Strength (VATS)..............................113

3.7.4 Test of previous knowledge of the target words 116

3.8 Procedure...117

3.9 Pilot Study and Rationale 123

3.9.1 Pilot participants 123

3.9.2 Piloting the massed and spaced treatments and their materials 124

3.9.3 Piloting the Vocabulary Achievement Test of Strength 125

3.9.4 Piloting the Questionnaires 126

3.10 Data Analysis...126

3.10.1 Analysis of Questionnaire data..126

3.10.2 Internal reliability of the questionnaire...127

3.10.3 Scoring the VST 129

3.10.4 Scoring the VATS...129

3.10.5 Interscorer and internal reliability of VATS...130

3.10.6 Statistical analysis 134

3.11 Summary 135

Chapter Four - Results...136

4.1 Introduction 136

4.2 RQ 1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?136

4.3 RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test? 141

4.4 RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength level between the immediate post-test and the delayed post-test? 146

4.5 RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed? 153

4.6 RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed learning?...159

4.7 RQ 5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test?...163

4.8 Further exploration of the data...163

4.8.1 The levels of strength of lexical knowledge as an implicational scale.......................163

4.8.2 Item analysis..171

Chapter Five - Discussion 176

5.1 RQ 1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?176

5.2 RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test? 181

5.3 RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength levels between the immediate post-test and a 4-week delayed post-test? 183

5.4 RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed? 185

5.5 RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed learning? 187

5.6 RQ 5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test? 189

Chapter Six - Conclusion 190

6.1 Introduction 190

6.2 Summary of findings 191

6.2.1 The effect of repetition learning with or without spacing on immediate acquisition of vocabulary knowledge at four strength levels 191

6.2.2 The effect of repetition learning with or without spacing on long-term retention of vocabulary knowledge at four strength levels 194

6.2.3 The effect of repetition learning with or without spacing on attrition between immediate and delayed tests of vocabulary knowledge at four strength levels 195

6.2.4 The spacing effect and parts of speech 197

6.2.5 Learners' perspectives on benefits for learning and engagement of the spacing effect 198

6.2.6 The effect of learners' perceived benefit for learning and engagement on the spacing effect 199

6.3 Limitations of the study 200

6.4 Implications for SLA and vocabulary learning theory 206

6.4.1 Productive and receptive vocabulary knowledge 206

6.4.2 Vocabulary learning rate 208

6.4.3 Part of speech effect on vocabulary learning 210

6.4.4 Number of repetitions and noticing 210

6.4.5 Spacing of repetitions 212

6.4.6 Student attitudes to massed/spaced learning 213

6.5 Implications for EFL pedagogy in Saudi Arabia and more generally 213

6.6 Future research directions 216

Bibliography 218

Appendices 233

Appendix A: Frequency profiles of vocabulary in the exercises using Lextutor BNC-20 234

Appendix B: learners' perception questionnaire (English) 234

Appendix C: learners’ perception questionnaire (Arabic) 237
Appendix D: Biodata and language learning history questionnaire (English) 239
Appendix E: Biodata and language learning history questionnaire (Arabic) 240
Appendix F: Translation Multiple-choice pre-test 241
Appendix G: Vocabulary Achievement Test of Strength (List A) 244
Appendix H: Vocabulary Achievement Test of Strength (List B) 249
Appendix I: Treatment Exercises 255
Appendix J: Participant Information Sheet and Consent Form.......................................268

List of Tables

Table 1.1 Vocabulary size tests used to measure Saudi EFL vocabulary size 4

Table 2.1 Summary of Bauer and Nation’s (1993) list of affixes (Milton, 2009, p. 104)............25

Table 2.2 What is involved in knowing a word (Nation, 2001, p. 27) 29

Table 2.3 Features of implicit and explicit forms of form-focused instruction (Graaff & Housen, 2009, p. 737)...37

Table 2.4 procedure of the study (Goossens et al, 2012, p. 968)..64

Table 2.5 Examples of vocabulary tests and the tested lexical aspects 74

Table 3.1 List of target words..107

Table 3.2 Procedure of the study 120

Table 3.3 KMO and Bartlett's Test..127

Table 3.4 Principal Component Analysis using Varimax with Kaiser Normalization rotation ..128

Table 3.5 Internal reliability of each questionnaire subscale 129

Table 3.6 Reliability of the achievement test items using Cronbach's alpha 131

Table 3.7 Reliability of the achievement test items by the split half method (odd-even items) using the Spearman-Brown coefficient 132

Table 4.1 Descriptive statistics for immediate tests scores on all strength levels 137

Table 4.2 GZLM analysis of Immediate Post-test scores...138

Table 4.3 Comparisons between massed and spaced conditions at each strength level: Immediate Post-test 140

Table 4.4 Descriptive statistics for delayed tests scores on all strength levels...........................142

Table 4.5 GZLM analysis of Delayed Post-test scores...143

Table 4.6 Comparisons between Massed and Spaced conditions at each strength level: Delayed Post-test. 145

Table 4.7 Overall GZLM analysis of immediate and delayed post-test scores, by learning condition and strength level...148

Table 4.8 Means for separate strength dimensions by time and learning condition...................151

Table 4.9 Overall GZLM analysis of immediate and delayed post-test scores, by learning condition and two strength dimensions 152

Table 4.10 Descriptive statistics for word class scores in the immediate post-tests at four strength levels by learning condition...154

Table 4.11 Descriptive statistics for word class scores in the delayed post-tests at four strength levels by learning condition...154

Table 4.12 Overall GZLM analysis of Immediate and Delayed Post-test scores, by Learning condition, Strength and POS 157

Table 4.13 Descriptive statistics for the questionnaire concerning learners' perceived difference between massed and spaced practice 160

Table 4.14 Tests of significant preference in responses to questionnaire items (Binomial test, omitting neutral responses)....161

Table 4.15 Student 7 in class 1 (immediate post-test, massed learning, word list B) 166

Table 4.16 Student 7 in class 1 (immediate post-test, massed learning, word list B), rearranged 167

Table 4.17 Imaginary data with added misfitting correct responses bold and misfitting incorrect responses in italics 168

Table 4.18 List A top and bottom scoring words at the receptive recall level (% correct) 171

Table 4.19 List B top and bottom scoring words at the receptive recall level (% correct) 172

Table 4.20 Descriptive statistics for the order of occurrence for each top and bottom scoring item in four exercises. 173

List of Figures

Figure 2.1 L2 components of vocabulary instruction (Laufer, 2005, p. 145)................................38
Figure 2.2 vocabulary profile of a typical learner (Meara, 1992/2010, p. 6) 51
Figure 2.3 typical forgetting curve (Schmitt, 2000, p. 131) 58
Figure 3.1 The typical experimental procedure for examining the spacing effect based on a single review opportunity. Adapted from Kang (2016, p. 13). 102
Figure 3.2 Design of the study using four practice tasks, immediate post-tests, and delayed post-tests. 102
Figure 3.3 Example of exercise one: receptive recognition task 108
Figure 3.4 Example of exercise two: productive recognition task 108
Figure 3.5 Example of exercise three: receptive recall task 109
Figure 3.6 Example of exercise four: productive recall task 109
Figure 3.7 Example of VATS Part 1 (productive recall)..114
Figure 3.8 Example of VATS Part 2 (receptive recall) 114
Figure 3.9 Example of VATS Part 3 (productive recognition) 115
Figure 3.10 Example of VATS Part 4 (receptive recognition)...115
Figure 3.11 Example of L2-L1 multiple-choice pre-test..117
Figure 4.1 Mean vocabulary scores at each strength level for Massed and Spaced conditions: Immediate post-tests. 139
Figure 4.2 Mean vocabulary scores at each strength level for Massed and Spaced conditions: delayed post-tests...144
Figure 4.3 Mean scores of immediate and delayed post-tests by learning condition and strength level (mean scores MAX=30)..147
Figure 4.4 Mean massed and spaced vocabulary scores at each strength level, and on the immediate and delayed post-test...149
Figure 4.5 Mean vocabulary scores for word classes by Massed and Spaced conditions over Immediate and Delayed post-tests (combined strength levels) 157
Figure 4.6 Mean vocabulary scores for Massed and Spaced conditions over Immediate and Delayed post-tests: each strength level separately..158
Figure 4.7 Mean ratings of students for perception of learning benefit (spaced high, massed low) 162
Figure 4.8 Mean ratings of students for perception of engagement benefit (spaced high, massed low) 162

Figure 6.1 Vocabulary teaching rate suggested for a theoretical course (Scholfield, 1991, p. 27) 209

Chapter One - Introduction

1.1 Research aim

Vocabulary learning is an essential component of mastering a foreign language. As such, vocabulary learning and instruction have become the focus of many studies in language research over the past five decades. Currently, it is widely accepted in research into vocabulary knowledge that vocabulary learning is incremental and that repeated exposures and recycling are necessary for vocabulary retention (Nation, 2001; Schmitt, 2010). However, there has been a marked neglect of research into how precisely the repetition and recycling of previously met words should be implemented in English as a foreign language (EFL) classrooms. A possible source of guidance into effective vocabulary repetition and recycling methods comes from one of the most robust findings in memory research in psychology, which suggests that spacing repetitions of whatever material is to be learned over a period of time with lengthy intervals between each repetition is better than massing repetitions in one lengthy session (Toppino & Bloom, 2002; Seabrook, Brown, and Solity, 2005). This finding has, however, barely been researched at all in the domain of the teaching and learning of foreign language vocabulary. The aim of this thesis is therefore to experimentally compare in classroom settings the relative effectiveness of spaced practice and massed practice on EFL vocabulary learning.

1.2 Background to the study

1.2.1 The problem of EFL vocabulary in the Saudi context

In the Kingdom of Saudi Arabia (KSA), English is considered as a key resource for public development due to its international prominence in many fields such as science, education, technology, politics, commerce, tourism etc. As such, learning English at state school in KSA is compulsory and starts from fourth grade of primary education, continuing through to school leaving after completing third grade of secondary education (i.e., nine years in all). Additionally, English is compulsory in many fields in higher education: most universities in KSA offer English as an undergraduate programme, and increasingly other majors are being taught through the medium of English (e.g., engineering, medicine). However, the outcomes of EFL classroom learning, especially in terms of vocabulary sizes attained, whether in general school education or higher education, have come under a lot of criticism in KSA for the past three decades (Al-Hazemi, 1993; Alsaif, 2011; Al Fotais, 2012).

Consequently, several studies have investigated possible reasons behind the low English proficiency among Saudi learners. The areas of investigation have included language learning strategies (e.g., McMullen, 2009; Alhaisoni, 2012; Javid, Al-thubaiti, & Uthman, 2013), learners' attitude and motivation (e.g., Abdul Haq & Samdi, 1996; Javid, Al-Asmari,& Faroog 2012), first language (L1) interference (e.g., El-Hibir & Al-Taha, 1992) foreign language anxiety (e.g., Abu-Gharabah, 1999; Al-Saraj, 2013; Alrabai, 2014), and indeed weakness in the instruction, textbooks and examinations (e.g., Al-Seghayer, 2015; Alenezi, 2016). These studies however mainly focused on language learning or teaching in general rather than learners' knowledge of specific aspects of language, such as vocabulary.

The importance of vocabulary in language learning and teaching is widely recognized by many scholars and researchers in the field of second/foreign language acquisition (e.g., Knight, 1994; Laufer & Nation, 1999; Nation, 2001; Schmitt et al, 2001; Read, 2004; Milton, 2009; Schmitt, 2010). However, vocabulary learning was almost overlooked in empirical research in Saudi Arabia until the early 90s. In particular, to the best of my knowledge, Al-Hazemi (1993) was the first study that investigated vocabulary knowledge of Saudi EFL learners. His study revealed that Saudi secondary school leavers and military cadets had a poor vocabulary size of around only 1,000 of the most frequent words in English. Al-Nujaidi (2003) reported a lower vocabulary knowledge of first year university students with an estimated vocabulary size between 400-700 words at the

2,000 and 3,000 most frequent word levels. Alsaif (2011) examined the vocabulary size of Saudi students at both school and university levels. Secondary school leavers knew on average 890 words, which is slightly below Al-Hazemi's (1993) estimate, while first year EFL major university students knew on average 2,452 words and final year EFL majors knew on average 3,252 words. Al Fotais (2012) found that second year EFL university majors had an estimated vocabulary size of 1,447 words of the most frequent words in English, which is lower than the estimated vocabulary size of first year university students in Alsaif (2011). Al-Masrai and Milton (2012) reported that first year EFL university students know on average between 2,000 and 3,000 words and final year EFL majors know approximately 5,000 words of the most frequent 10,000 words of English.

The variation in vocabulary size estimates in these studies might be in part due to the different tests that were used to assess vocabulary size of Saudi EFL learners. Different vocabulary size tests come in different formats and usually measure distinct aspects of vocabulary knowledge (see Table 1.1). For instance, the X-Lex (Meara & Milton, 2003) uses the *lemma* as a unit of measurement (i.e., counting a base word and all its inflections as single units), while the Receptive Vocabulary Test (RVT) (Alharthi, 2012) uses the *word family* as a unit of measurement (i.e., counting a base word and all its derivations and inflections) (see Section 2.3). As such, a test that uses the lemma as a unit of measurement may produce a higher vocabulary size score than a test that uses the word family as a unit of measurement. According to Milton (2009, p. 12), "to compare a vocabulary size measurement made using word families with one made using lemmas, multiply the score in word families by 1.6 to get a rough (very rough) equivalent score in lemmas". Therefore, the vocabulary size score in Al Fotais (2012) would roughly be 2,315 lemmas, which suggests that the findings of the Alsaif (2011) and Al Fotais (2012) are not really conflicting.

Furthermore, it is arguably understandable for speed and simplicity why many studies use a yes/no test format, such as Eurocentres Vocabulary Size Test (Meara & Buxton, 1986). However, this type of tests only measures if learners know that a word exists as a form in English and does not measure if the learners know the word's meaning. Evidently, if a learner knows that, for example, *track* exists as an English word without knowing what it means, that learner will get a mark for that on a yes/no test, but not in a multiple-choice test where the word that is most similar in meaning has to be selected from a range of alternatives. For this reason, therefore, some vocabulary size measurements might have inflated the vocabulary size estimates of Saudi EFL learners, which makes the low figures even more worrying.

Study	Test	Format	Aspect of receptive vocabulary knowledge
Al-Hazemi (1993)	Eurocentres Vocabulary Size Test (Meara & Buxton, 1986)	yes/no	knowledge of written form
Al-Nujaidi (2003)	Vocabulary Levels Test (Nation, 1990)	multiple-choice	knowledge of the form-meaning link
Alsaif (2011)	X-Lex (Meara & Milton, 2003)	yes/no	knowledge of written form
Al Fotais (2012)	Receptive Vocabulary Test (Alharthi, 2012)	multiple-choice	knowledge of the form-meaning link
Al-Masrai and Milton (2012)	XK_Lex (Al-Masrai, 2009)	yes/no	knowledge of written form

Table 1.1 Vocabulary size tests used to measure Saudi EFL vocabulary size

Research into the relationship between vocabulary size and language proficiency suggests that learners need to know, on average, between 8,000 and 9,000 words to comprehend 98% of non-specialist authentic text (Nation, 2006), which might be regarded as a prerequisite ability for progressing to discipline specific academic texts at university. A vocabulary size of as much as 20,000 words may be needed to read academic texts (Nation & Webb, 2011). In order to reach adequate comprehension in speaking and listening, which is less challenging than reading, according to Milton (2009, p. 59), beside knowledge of "the most predictable and formulaic exchanges", EFL learners need to know at least 3,000 words out of the most frequent 5,000 words in English. Accordingly, all the earlier estimates of Saudi EFL learners' vocabulary size indicate that there is a vocabulary learning deficiency, which demands further investigation to examine some possible reasons and solutions.

Such findings are not of course only a feature of Saudi Arabia. Many EFL countries are highly likely in a similar position of students at tertiary level failing to reach a desired English proficiency level suitable for genuine independent language use and English medium study. Examples are Thailand and Taiwan. In the latter, for example Chiang (2018) found a mean vocabulary size only a little above 2,000 in first year technical university students. Hence any light that the present study is able to throw on such a situation has wider interest than just for Saudi Arabia.

Several studies, mostly dissertations, have examined some of the possible reasons behind poor vocabulary knowledge among Saudi EFL learners. Different investigations revealed factors including: insufficient use of vocabulary learning strategies (Al-Shuwairekh, 2001; Alyami, 2011; Alhatmi, 2012), ineffective vocabulary instruction in the classroom (Alhaidari, 2006), poor vocabulary input in textbooks (Al Fotais, 2012; Alenezi, 2016), limited vocabulary knowledge of English language teachers, lack of vocabulary exposure outside the classroom (Alsaif, 2011) and insufficient vocabulary recycling in textbooks (Al Fotais, 2012).

Furthermore, some studies suggested a number of recommendations for tackling the issue of poor vocabulary knowledge among Saudi EFL learners, such as training learners to be independent and autonomous (Al-Shuwairekh, 2001), training learners in effective vocabulary learning strategies (Alhatmi, 2012), encouraging learners towards extracurricular activities involving English (Alyami, 2011), increasing vocabulary input in textbooks (Alsaif, 2011; Alenezi, 2016), regular short-term testing of vocabulary knowledge (Alharthi, 2012), and systematically recycling vocabulary in textbooks (Alsaif & Milton, 2012; Al Fotais, 2012).

It should be noted that the term *repetition* is often used particularly of what the learner deliberately and consciously does himself, whether prompted by a teacher or textbook or not, and is often short term (i.e., massed). *Recycling* is more often used to refer to repetition engineered by the teacher or textbook or reading materials, may be undetected by the student, and is often over a longer term, so more spaced (see further Section 2.6.5). From the above, it can be inferred that many aspects of repetition and/or recycling have in fact been indirectly researched, but not precisely the massed versus spaced dimension. For example, repetition strategies of various types loom large among vocabulary learning strategies. While these repetition strategies include issues like what unit is repeated (e.g., word, phrase etc.), what information about the unit is repeated (e.g., word alone, or word and L1 translation, or word and example sentence), in what mode (e.g., speaking, writing, reading, or hearing repeatedly), and the like, they do not however examine the timing of the repetitions. Furthermore, only the repetition of word lists of some sort is usually in focus, not repetition in textbook-style exercises, which is the focus of this study.

Extracurricular reading or listening activities, mentioned above, will inevitably mean that common English words are met repeatedly, but these encounters will very likely be randomly spaced. Testing is a way of repeating information that has already been introduced which has been seen as having not only usefulness to measure what has been learned but also a special value for learning itself (Roediger & Butler, 2011), but the timing of testing has received less attention. Finally, better recycling of new vocabulary in textbooks, not only within the unit where it first appears, is yet another special form of repetition, ultimately supported by ideas such as those of Ebbinghaus (1885/1993) about the decay rate of memory (Rubin, Hinton, and Wenzel, 1999). Thus, none of the above precisely target the massed-spaced distinction.

Additionally, there have been a few studies of vocabulary teaching in Saudi Arabia which refer to repetition or recycling in some sense (e.g., Al Akloby, 2001; Alghamdi, 2013). However, once again these studies do not address the present study's concern with the timing of such repetition or recycling but rather with more general issues of the kind of practice tasks that are used, each of which typically involves some repetition/recycling of learned material. For instance, getting students to say words aloud repeatedly as soon as they are presented, giving them weekly vocabulary quizzes, or making sure that they meet new words in multiple types of exercise are all forms of repetition but might well differ considerably in their effects due to the differing nature of the repetition task itself. In the present study this variable is carefully controlled by varying the precise type of task systematically based only on four aspects of word knowledge that are targeted, along with spacing.

In sum, no studies have been carried out on vocabulary repetition of the specific types which I am concerned with in Saudi EFL classroom settings, and with attention to its timing.

1.2.2 Gaps in the study of repetition/recycling in foreign language vocabulary learning and teaching in general

The dearth of research attention to the timing of teacher-led, conscious, and vocabulary-oriented repetition or recycling in common classroom EFL exercises is not only to be found in Saudi Arabia, however. The need for repetition to maintain and extend vocabulary knowledge is probably one of the most consistent findings from research into vocabulary learning conducted in many countries (Nation, 2001; Schmitt, 2010). However, the scheduling of repetitions, which might be equally important and could have an impact on long-term vocabulary retention, has been neglected in language research. Nevertheless, a great deal of psychological research into the learning of many kinds of information supports the benefit of spacing. Based on a robust finding in memory research stemming from Ebbinghaus (1885/1913), learners have been found to retain information better when instruction/learning sessions are repeated over a period of time with lengthy intervals between each learning session, as opposed to learning information in a single lengthy learning session (e.g., Melton, 1970; Dempster, 1987; Anderson, 1990; Dempster & Farris, 1990; Dempster, 1991; Russo, Mammarella, & Avons, 2002; Toppino & Bloom, 2002; Seabrook et al., 2005). In an attempt, therefore, to bridge the gap between psychological research and applied linguistic vocabulary research, this thesis compares the relative effectiveness of spaced practice and massed practice on vocabulary learning through simple practice exercises in an authentic EFL classroom setting.

A review of the relevant literature (see Chapter 2) reveals that in fact only three studies have examined the spacing effect and its impact on vocabulary learning in authentic classroom settings, namely, Sobel, Cepeda, and Kapler (2011), Goossens, Camp, Verkoeijen, Tabbers, Bouwmeester, and Zwaan (2012), and Schuetze (2015). It should be noted that Sobel et al. (2011) investigated the effect of spaced repetitions on L1 English vocabulary learning by young children, while only Goossens et al. (2012) and Schuetze (2015) investigated the effect of spaced repetitions on L2 vocabulary learning. Sobel at el. (2011) examined the difference between spaced repetition and massed repetition on vocabulary retention among fifth-grade primary school students in an Ontario middle school. The study reported that retention of spaced vocabulary was three times higher than retention of massed vocabulary. Goossens et al. (2012) examined the difference between spacing and massing vocabulary in learning sessions among third-grade L1 Dutch primary school students learning EFL. The results indicated better performance for vocabulary learned in the spaced condition. Finally, Schuetze (2015) compared the impact of two types of spacing (i.e., equal distribution and expanding distribution of spacing). In two studies, the two spacing methods were examined in learning vocabulary of German as a foreign language at university level. The results of these studies did not reveal a statistically significant difference between the two spacing methods. However, there was an increase in the retention score in the second study due to increasing the number of repetitions from three to four repetitions.

There are, however, some methodological shortcomings of the previous studies that will be addressed in the current study. First, the ecological validity of Sobel et al. (2011) and Schuetze (2015) suffers from the fact that only one type of learning task was used, which does not reflect what often happens in a real classroom learning session. Usually, a vocabulary teaching session involves students doing several different tasks/exercises. The current study, therefore, introduces four task types, each associated with a different type of knowledge of the target words being learned (see point six below). Second, pre-existing knowledge of the target items was not assessed in these previous studies. For example, Schuetze (2015) only estimated previous knowledge of the items based on a questionnaire that was used to single out participants who had experience with German. Sobel et al. (2011) however decided that the items were unknown/new to the target participants based solely on the researchers' judgment, while Goossens et al. (2012) did not examine prior knowledge of the target items at all. It seems imperative in any similar line of research on vocabulary retention to establish a baseline; thus, examining previous knowledge of the target items was not ignored in the present study.

Third, the number of items selected in the previous studies was probably insufficient to ensure reliability and so justify further statistical analysis. For example, Sobel et al (2011) used only four words. The present study uses far more.

Fourth, the selection of the target items in the previous studies did not consider the possible effect of word class on vocabulary learning and/or retention. For example, Sobel et al. (2011) selected four adjectives, two nouns and two verbs, while Goossens et al (2012) selected 25 nouns. Word class however may affect the ease or difficulty of vocabulary learning and it might be "prudent to control for word class in all vocabulary research" (Schmitt, 2010, p. 160). This could be accomplished by targeting one part of speech, as Goossens et al. (2012) did, or by choosing an equal number of words from different word classes, as Schuetze (2015) did, but in larger numbers. Fifth, word length was another factor that was not addressed in previous studies. Schuetze (2015) pointed out that longer and phonologically similar words were learned the least. However, the small number of long words used in his study did not allow for a statistical analysis. In the present study length was controlled by making it uniform across all words.

Sixth and finally, the instruments that have been used to assess vocabulary retention in the previous studies might not have been fully sufficient to capture the effect of spaced practice on vocabulary learning since they did not take into account that learning a word is not a unitary event. Rather, words have a number of aspects to be learned (i.e., types of lexical information) and success in learning one does not guarantee success in learning another (i.e., these aspects may be regarded as levels of word knowledge which is each associated with a different learning task and test type). For example, the above three studies all used variants of a receptive recall task, where the meaning of a target word form that is supplied has to be provided (through paraphrase or translation). Each of these studies used only one type of measurement that assessed vocabulary knowledge at a quite demanding level, since no multiple choices were offered (see Laufer & Goldstein, 2004). It is generally recommended however that research into vocabulary learning uses multiple measures to provide a more comprehensive picture of vocabulary knowledge and learning (Nation, 2007; Milton, 2009). This could be achieved by examining "receptive/productive mastery, different types of word knowledge, degree of mastery of an individual word knowledge aspect, contexts of use, etc., or some combination of these" (Schmitt, 2010, p. 22). The current study, therefore, selected four such aspects/levels to investigate.

From the above, then, it may be seen that research about the effectiveness of spaced practice on vocabulary learning in EFL classrooms anywhere is markedly limited, not only in Saudi Arabia. Hence, the current study has considerable interest for FL vocabulary learning/teaching researchers in general, aside from teachers and researchers in Saudi Arabia. Results from Sobel et al. (2011), Goossens et al. (2012), and Schuetze (2015) are suggestive of the value of further examination of the phenomenon of the spacing effect and the role it could play in vocabulary learning in the EFL classroom. However, the investigation of this effect is clearly in need of further attention. Consequently, it is hoped that the current study will make a valuable contribution to the field of learning and teaching of English as a foreign language in Saudi Arabia, in particular, and to the field of foreign and second language vocabulary acquisition in general.

1.3 Methodology

The experiment conducted in this thesis to compare between the effectiveness of massed practice and spaced practice in instructed EFL vocabulary learning will be described in full in Chapter 3 but is briefly introduced here. The present study took place in two intact classrooms of first year EFL university students. The study was designed to fit into normal classroom time over a period of 10 weeks. The study first established a baseline of vocabulary knowledge by testing existing knowledge of the target items, then introduced the learners to those items through one of two types of treatment. In the massed practice treatment, the target words were practiced four times in one learning session, where each practice task targeted a different level of vocabulary knowledge of the same words. In the spaced practice treatment, a parallel set of target words was practiced once every week for four weeks, with each week targeting a different level of vocabulary knowledge. This teaching phase was followed by immediate post-tests to measure vocabulary learning at each level of vocabulary knowledge at peak attainment, and delayed post-tests which measured retention of the target words after four weeks.

The participants selected for the study were Saudi EFL university students for two reasons. First, previous studies that examined the effect of spaced practice on vocabulary learning in actual classroom settings were conducted in primary school classrooms. As such, the current study hoped to shed light on this issue at university level. Second, because I hold a position as a lecturer at Taif University in KSA, it was feasible to get permission for the study and to teach two intact classrooms of students without interrupting their actual university study programme.

1.4 Delimitation of the research

In order to keep the study manageable, I only introduced and had the students practice a limited range of lexical information about each of the target words. Hence, the two vocabulary repetition timings examined in this thesis are only claimed to affect learning of receptive and productive vocabulary knowledge of the written form and one meaning of each word, in both recognition and recall modes. As such, this thesis does not examine or make suggestions for learning other aspects of vocabulary knowledge such as associations, grammatical functions, and collocations.

In addition, this thesis does not itself attempt to test or investigate any theoretical basis for the spacing effect. However, a review of the various theories that attempt to explain the spacing effect is provided in Section 2.4.1.

1.5 Thesis outline

This thesis is divided into six chapters. This first chapter presented the aims of the study, a brief review of the spacing effect, a brief discussion about previous research on the spacing effect in vocabulary learning classrooms, and the delimitation of the study. Chapter Two surveys the relevant research literature to provide a background for the research problem. Chapter Three provides a description of the method and presents the pilot study which was employed to trial the research tools and procedures intended to be followed in the main study. Chapter Four presents the results of the study in relation to each Research Question (RQ). Chapter Five presents a discussion of the study’s findings in relation to each RQ. Finally, Chapter Six discusses major findings, examines how these findings could inform EFL vocabulary learning, discusses the limitations of the study and makes suggestions for pedagogy and future research.

Chapter Two - Literature Review

2.1 Introduction

The current thesis aims at investigating the effect of massed practice and spaced practice on vocabulary learning among Saudi EFL learners at university level, at four different levels of word knowledge. Therefore, this chapter will discuss issues related to the main three themes of this thesis, namely, the nature of vocabulary and vocabulary learning, the spacing effect, and vocabulary testing.

This chapter will first discuss the importance of vocabulary in language learning, the meaning of *word*, what knowing a word involves, and how many words EFL learners need in order to speak, write, and read English sufficiently. It will then discuss research related to vocabulary learning. This discussion will focus on incidental and intentional vocabulary learning. Then, further discussion will be provided in terms of identifying a few factors that might affect vocabulary learning in the present study.

This chapter will proceed to review the phenomenon of, and the main theoretical explanations for, the spacing effect. Then, a review of laboratory-based and classroom-based research on the spacing effect and its impact on vocabulary learning will offer corroborating evidence for the benefit of spaced practice over massed practice.

In the final part of this chapter, a discussion on vocabulary testing will be provided. This discussion will include the importance of vocabulary testing, multiple measures in vocabulary testing, current vocabulary tests, some methodological considerations when choosing vocabulary tests, and the rationale behind the selection of the type of measurements used in the current study. The chapter will end by providing a summary of the chapter and a discussion of some methodological considerations regarding real-world classroom-based studies. Following the summary of the chapter, the research questions/hypotheses of this thesis will be presented.

2.2 The importance of vocabulary learning in foreign language learning

Vocabulary is fundamental to language learning because “language ability is to quite a large extent a function of vocabulary size” (Alderson, 2005, p. 88). As Milton (2009, p. 3) succinctly puts it, “[w]ords are the building blocks of language and without them there is no language”. Prior to the 1980s, however, the importance of vocabulary in foreign language learning was not recognised as “the words *lexis* and *vocabulary* [were] remarkable by their absence from either chapter headings or indexes in the major writers on syllabus” (O’Dell, 1997: p.258).

According to Milton (2009), there are three main reasons behind the lack of attention to vocabulary learning, testing, and teaching in much of the literature on SLA during the last half century. First, traditional foreign/second language teaching methodologies (e.g., Grammar Translation Method, Audio-lingual Method) view vocabulary as far from being of pivotal importance in language learning. Commonly, the emphasis in these learning methods is on how learners acquire rules and systems of a language, especially its grammar. Indeed, theoretical linguists of the time also contributed to this with their extreme focus on syntax, and to an extent phonology, with little attention to the lexicon in language. There was a belief that, despite vocabulary being subject to the rules and systems of language, these rules are independently developed “regardless of which words, or how many words, were being used to form them” (Milton, 2009, p. 1). Seal (1991) even claimed there were periods when learning too much vocabulary “was regarded as a positively dangerous thing” (p. 296). As such, it is common in these learning methods, at least at the early stages of language learning, that vocabulary input is intentionally reduced to only serve as a medium of teaching language rules or to aid in motivating learners. According to Milton (2009), the effect of learning methodologies that promoted such structural linguistic approaches to language learning was so robust that it extended to later learning methodologies, such as the Notional-functional approach and communicative approaches in the UK, where the importance of vocabulary should have been more pronounced.

The second reason is a product of a widespread belief among many foreign language teachers and learners that language proficiency can be achieved with a limited vocabulary size. This belief could be attributed to research that predated modern corpus analysis. According to Milton (2009) it is still widely believed among some teachers and textbook designers that learning a set of 850 words from Ogden's (1930) *Basic English* is sufficient to learn the modern English language. In fact, that kind of approach is concerned only with how many words people need in order to express themselves (i.e., for production) where indeed, with some difficulty and much use of paraphrase, it is possible for a person to say or write a great deal using only a small number of words. However, this does not extend to reception, especially reading, where the speaker cannot control the number of different words that appear in the input that he/she receives. Here, based on data from the British National Corpus (BNC), Nation (2006) has suggested that language learners need to know around 9,000-word families in order to read authentic non-specialist texts with reasonable ease (e.g., novels, newspapers, magazine articles) (see Section 2.3 for a description of the different methods of counting words).

The third reason stems from another belief, that a large amount of vocabulary is retained through learning which is unintentional (Ellis, 1994), and a limited amount of vocabulary is retained when learning is intentional (Harris & Snow, 2004). Thus, vocabulary teaching is viewed as a waste of time and effort because it is supposed that learners will eventually learn more vocabulary simply by unintentional exposure to the language. Such a view on the dominant effect of extensive comprehensible input is particularly associated with Krashen (e.g., 1985) and his Natural approach to language learning and teaching. However, actual research suggests that "the vocabulary uptake from truly incidental language exposure is usually negligible and that successful learners acquire large volumes of vocabulary from the words explicitly taught in the classroom and supplement their learning by targeting vocabulary in activities, like learning the words of songs, outside of class" (Milton, 2009, p. 2).

Finally, yet another false belief that might be mentioned here was the idea that existed for a long time that if only students were taught proper guessing skills they did not need to learn a lot of vocabulary. However, it is now known that the meaning of most words is not guessable as there are insufficient clues in natural language text and discourse as even native speakers and expert guessers cannot usually do better than 50% of guessing the meaning (e.g., Laufer, 1997; Bornmann and Munby, 2005).

However, despite earlier misinformation about the importance of learning vocabulary in foreign language learning, attention to vocabulary issues grew steadily over the past four decades and now there is a substantial body of research on vocabulary learning (e.g., Read, 2000; Nation, 2001, 2011; Schmitt, 2000, 2008, 2010; Gu, 2003; Laufer, 2009; Milton, 2009; Folse, 2011). The investigation of vocabulary learning has been carried out in many areas, such as the role of word frequency in vocabulary learning (e.g., Coxhead & Nation, 2001), vocabulary assessment (e.g., Laufer, Elder, Hill, & Congdon, 2004; Nation, 2006; Read, 2007; Milton, 2009, Schmitt, 2010), vocabulary repetition (e.g., Rott, 1999; Nation, 2001; Webb, 2007), explicit and implicit vocabulary learning (e.g., Ellis, 1994; Gass, 1999; Laufer & Hulstijn, 2001).

Evidence from vocabulary research suggests that there is a positive correlation between how many words learners know and how well they perform in different language skills (i.e., listening, reading, speaking, and writing). In a comprehensive study, Milton, Wade and Hopkins (2010) examined the relationship between the scores of 30 intermediate and advanced level learners on two vocabulary size tests (i.e., X-Lex, Meara & Milton, 2003; A-Lex, Milton & Hopkins, 2006) and their scores in the International English Language Testing System (IELTS). Results indicted a positive correlation between the learners' orthographic vocabulary size (i.e., X-Lex scores) and the learners' reading and writing skill level, as well as, a strong correlation between the learners' phonological vocabulary size (i.e., A-Lex scores) and the learners' listening and speaking skills.

2.3 What does *word* mean?

This thesis aims to investigate the spacing effect on vocabulary learning among Saudi EFL learners at university level. Since *vocabulary* is widely defined as comprising the words of a language (Summers & Stock, 1992), defining what the term *word* means is therefore necessary before examining any vocabulary learning methods or vocabulary tests.

A check of the definition of *word* in online dictionaries reveals that there are many different definitions. For example, in *Oxford Online Dictionary* (2017), a *word* is "[a] single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed"[1]. In Merriam-Webster Dictionary (2017) a word is "a speech sound or series of speech sounds that symbolizes and communicates a meaning usually without being divisible into smaller units capable of independent use"[2]. Cambridge Dictionary (2017) describes a *word* as "a single unit of language that has meaning and can be spoken or written"[3].

Based on the definitions above, it may seem that defining a *word* is straightforward. A word is basically a unit of language consisting of letters (or sounds) with an empty space on either side, that convey meaning or function in a spoken or written form. Such a definition is generally acceptable to count, for example, how many *words* a student is writing in an essay. However, adhering to these simplistic descriptions of the term *word* would be problematic in vocabulary

[1] Word. (2017). In Oxford Online Dictionary. Retrieved from https://en.oxforddictionaries.com/definition/word
[2] Word. (2017). In Merriam-Webster Dictionary. Retrieved from https://www.merriam-webster.com/dictionary/word
[3] Word. (2017). In Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org/dictionary/english/word

research. There are various methods of counting words and using each counting method will lead to different results (Schmitt, 2010). For example, the words *rings, rang, rung, ringing,* and *ringed* are affixed variations of the word *ring*. Therefore, are these forms counted as a single word or several words? How about *ring rings* as a noun? Is that a separate word or not from the verb? And indeed, as a noun is the *ring* in *There was a ring on her finger* a separate word from *ring* as in *There was a ring at the door* where the meaning is quite different? Yet further dilemmas arise, this time of a sequential nature, from the conflict of the written space criterion with the idea that a word is a single unit of meaning. Do we regard items like *pull my leg* 'play a joke', *put up with* 'tolerate' and *of course* 'definitely' as each made of multiple words due to the spaces or as single words due to the unitary meaning? The reverse problem arises for contracted forms like *we'll* and acronyms like *KFC*. Although written with no spaces, are they in fact sequences of more than one word?

The issue of not having a clear concept of what is meant by a *word* can be observed in earlier tests and estimates of native speaker vocabulary size. According to Goulden, Nation, and Read (1990, p. 356), early vocabulary size estimates that suggested native English speakers know around “216,000 words (Diller, 1978) or 80,000 words (Miller and Gildea, 1987) [were] clearly inflated”. Some of the possible reasons behind such misleading vocabulary size estimations are attributed to methodological flaws that failed to answer the following questions: *what is meant by a word? how to sample a wordlist to make a vocabulary test?* and *how to test vocabulary?* (Goulden, et al., 1990; Nation, 1993; Nation & Waring, 1997). In fact, well-designed research that counted words systematically using frequency information suggested that vocabulary size estimates of well-educated native speakers may vary on a range from 10,000-word families (Milton & Treffers-Deller, 2013) to 17,000-word families (Goulden et al., 1990; Schmitt, 2010).

Dilemmas about what is a *word* also spill over into research on vocabulary learning and teaching. In the present study, it is necessary to be clear what kind of word-like entities will be taught, learned, and tested so as to be able to say which learning condition (i.e., massed or spaced) leads to learning of more words in a clearly specified way. Consequently, it is essential to establish a clear methodology of defining and counting words. In general, there are four main methods of doing this, namely defining words in terms of *tokens, types, lemmas,* or *word families*. The following describes each of these units.

2.3.1 Tokens and types

The terms *token* and *type* are very useful in vocabulary research, especially in corpus research (Schmitt, 2010). The term *token*, frequently referred to in dictionaries as a *running word,* is used to describe the method of counting every word occurring in written or spoken texts, while the term *type* is used to count the number of different words. Both usually take a word to be whatever is sequentially delimited by a written space. For example, there are nine tokens in a sentence like ‘The boy kicked the ball to the other boy’, and six in 'I cannot put up with this’. This method of counting words is useful to determine how many words in a text, for example, counting the number of words in a student’s essay assignment. In learning and teaching, however, although what the learner hears, and reads are always word tokens, what we imagine them actually learning (storing in memory) is words in some higher-level sense, at the very least types.

In the first example above, if the goal is to count types, then only the number of *different words* is considered, so there are six types: *the boy kicked ball to other*. This method of counting is more useful in teaching and assessing vocabulary knowledge among language learners, since when we speak of students learning words we usually mean different words, not repetitions of the same word. The definition of a type, however, would treat occurrences of *kicks* and *kicked* as two different types, also *boy* and *boys,* which, as will be discussed later, is not necessarily desired for pedagogical purposes.

Even with these seemingly simple approaches, there are some issues. One is how to deal with sequential problems in speech, where there are no 'spaces' to help (Milton, 2009). This includes contracted forms, which are more common in speech than in writing (e.g., *I'm, I'll, gotcha!*). The decision is often made to treat contracted forms as sequences of tokens or types. Another decision often made is to exclude certain forms as having insufficient meaning to be considered. In speech, for instance, it is not clear how to count forms that are merely pause fillers (e.g., *um, uh, er, ah*), and there has been a general linguistic debate over whether proper names really have meaning (e.g., Gabriel, 1990) or are worth counting for pedagogical purposes. In response to these issues, there is now a general consensus among researchers that "numbers, proper nouns and names, and false starts and mistakes [should be] excluded from word counts" (Milton, 2009, p. 9).

Overall, counting the number of words in type units produces large numbers of 'different' words due to inflected forms being treated as separate types: for instance, an English verb has at least four type forms (*kick, kicks, kicked, kicking*). Some linguistic theories provide entries for every such type in their lexicons (e.g., Lexical Functional Grammar), but there is conflicting psycholinguistic evidence over whether native speakers store morphologically complex words (e.g., *kick+ed*) separately from related simple forms (e.g., *kick*) in the mental lexicon or not (Leminen, Lehtonen, Bozic, and Clahsen, 2016) (see further next section).

In teaching and learning, there are some instances where a textbook or teacher might present inflected forms separately to be learned (i.e., types). At elementary levels, the meaning of *kicked* might be explained as a separate item from *kick*, and this would be even more likely for irregular inflected forms like *sang* or *went*. However, this is not the usual practice at later levels of proficiency such as that which the current study is concerned with. Teachers and textbooks present, list, count and test *kick* as a word in a sense that includes all its inflected forms, and it is supposed that if the type *kick* is explained, the learners will know the meaning of *kicks, kicked,* and *kicking* without further separate attention to those words. Hence for the present research a higher-level definition of 'word' is used (i.e., lemma), which gives smaller and more pedagogically relevant estimates of vocabulary knowledge (Milton & Treffers-Daller, 2013).

2.3.2 Lemmas and word families

It is reasonable for language learners who are hopeful of achieving native-like proficiency to aim at building a vocabulary size comparable to the vocabulary size of native speakers. However, learners would be overwhelmed to discover that vocabulary size estimates of native speakers, based on counting *tokens* and *types*, suggest that secondary school native speakers acquire 20,000 words every year and reach up to 200,000 words by the end of secondary education (Diller, 1978; citied in Goulden et al., 1990). These early inflated estimates are due to the methods that were employed in the word counts (Nation, 1993).

There are two additional methods of defining and counting words which treat certain groups of types together as one 'word', and offer smaller and, for most purposes, more valid estimates of vocabulary knowledge. The first method is called *lemmatisation* and it involves counting "a headword and its most frequent inflections [as a single unit], and this process must not involve changing the part of speech from that of the headword" (Milton, 2009, p. 10). For instance, the word *heat* as a verb is the lemma for *heats*, *heating*, and *heated*. *Heater* and *heaters* are not part of the lemma *heat* because they are nouns, not verbs. Instead, *heater* and its plural inflection *heaters* are combined to form another separate lemma. Similarly, the noun *heat* is a separate lemma from the verb *heat*. This is in effect the definition of 'word' used by most dictionaries: each main entry describes one lemma rather than one type. It is also what is most often meant when teachers or researchers refer to teaching and learning 'words'.

Schmitt (2010) makes some valuable additional points about the psycholinguistic justification for *lemmas*. First, some psycholinguistic studies support the lemma concept by suggesting that, rather like a traditional dictionary, regular nouns and verbs are stored in the mind by native speakers as lemmas in their root form, and inflected forms are only created as suffixes that are added later in the process of production (Aitchison, 2012), or stripped off in reception. In the aforementioned view, the mind would not store inflected types separately, but only store the base form of a lemma while processes of production and reception deal with its inflected types when a speaker needs them.

Second, it is important to note, however, that there are around 200 irregular past tense verb types and many partially irregular plural nouns in English (Schmitt & Marsden, 2006) which are often seen by psycholinguists as treated differently. Each irregular type would be stored individually as a separate lemma rather than generated from the base, because there is no general process that can form irregular words from their base.

Third, the implication for learning is that learning lemmas is less work than learning types, as there are far fewer lemmas (Schmitt, 2010). For example, if learners know that adding '*-s*' to the end of a regular singular noun (e.g., *book, desk)* transforms it into a plural noun (e.g., *books, desks*), and adding '*-d*' to the end of a regular verb (e.g., *play, walk*) changes it to the past tense (e.g., *played, walked*), then learning these inflected forms should be quite easy. However, it should be noted, that learners must memorize the different forms of irregular words separately. For example, a learner's ability to connect forms of nouns with irregular plural forms (e.g., *goose*: *geese, sheep*: *sheep*) and verbs with irregular past forms (e.g., *eat*: *ate, lose*: *lost*) should not be assumed, yet usually these types would be regarded as included within one lemma in each case. As such, the present study works with the notion of lemma as the unit that is taught, learnt, tested, and counted.

The last and highest-level method of defining and counting words involves treating different word types, beyond the inflected forms, provided they are semantically as well as formally related as a single *word family*. The term *word family* is defined as "a base [or root] word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (Bauer & Nation, 1993, p. 253). Thus, words from different word classes, and with derivational as well as inflectional affixes, may belong to the same word family. For instance, the following words are formed around the base form *heat* (noun) and belong to the same *word family*: *heat (verb), heats, heated, heating, heater, heaters, unheated, preheat, reheating* etc. Thus, several lemmas, each containing a number of types, are united in one family.

This method of counting words naturally generates even lower estimates of vocabulary size than counting words as lemmas. For example, the vocabulary size of well-educated native speakers may vary between 10,000-word families (Milton & Treffers-Deller, 2013) to 17,000-word families (Goulden et al., 1990; Schmitt, 2010). There are, however, a number of implications of using the *word family* as a word unit in vocabulary testing, teaching, and learning. In vocabulary testing, choosing words from a few thousand-word families is easier to do in a systematic way than choosing words from tens of thousands of word lemmas in a dictionary (Milton, 2009). It must be noted, however, that a word family that as a whole is frequent may contain some quite infrequent individual word types. Consequently, in vocabulary learning, it is less clear than with lemmas that a non-beginner who knows one member of the set knows all. It is reasonable to assume that if we have taught *act*, and students have learned it, they also know *acts* and *acting* (i.e., members of the lemma *act*), but it seems a less justifiable claim that the students also will know, without further work, *activation, reactivity* and *unactionable*, which are all part of the word family.

Furthermore, it should be noted that there are no clear rules about many details of what must or must not be included even under a *lemma*, let alone a *word family*. With respect to what affixes to include, Milton (2009) suggests that it is possible to determine if a form with a derivational or inflectional affix is under the same *lemma* or the same *word family* based on the seven affix levels classified by Bauer and Nation (1993) (see Table 2.1). In this view, *lemmas* can be defined by including word types that include affixes from the three most frequent levels.

Level	***Affix***
1	n/a different form is a different word
2	Regular inflections: plural, 3rd person singular present tense, past tense, past participle, -ing, comparative, superlative, possessive
3	-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un- (all with restricted uses)
4	-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in- (all with restricted uses)
5	-age, -al, -ally, -an, -ance, -ant, -ary, -atory, -dom, -eer, -en, -ence, -ent, -ery, -ese, -eque, -ette, -hood, -i, -ian, -ite, -let, -ling, -ly, -most, -ory, anti-, ante-, arch-, bi-, circum-, counter-, en-, ex-, fore-, hyper-, inter-, mid-, mis-, neo-, post-, pro-, semi-, sub-, un-
6	-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-
7	Classical roots and affixes

Table 2.1 Summary of Bauer and Nation's (1993) list of affixes (Milton, 2009, p. 104)

On the other hand, in this approach, a *word family* can be defined by selecting derived and inflected forms that use affixes from the six most frequent levels. However, this raises some difficulties. If applied simplistically it would put *hospital* and *hospitality* in the same word family (since *-ity* is level 4) and *fruit* and *fruition* in the same word family (since *-ion* is level 6). Clearly semantic connection has to be used as a criterion as well but deciding what is or is not semantically related is often difficult (Brown, 2017). Furthermore, those who determine word families often seem more influenced by sameness of form than connection of meaning. The *familizer* tool on the Compleat

Lexical Tutor (Cobb, 2019)[4] for instance puts *act* and *action* in the same family but *decide* and *decision* in two different families.

In sum, choosing a suitable unit of measuring vocabulary depends on the purpose, participants, and resources of the study. As mentioned earlier, due to different methods of word counting, there is a large discrepancy in vocabulary size estimates of native speakers in the literature. To compare vocabulary size estimates between a study that used the *lemma* and another that used the *word family*, Milton (2009) suggests multiplying the estimate in *word families* by 1.6 to get an approximate equivalent in *lemmas*.

In most cases, increases in general proficiency levels of EFL learners lead to increases in their derivational knowledge (Schmitt & Zimmerman, 2002). Therefore, *types* are more likely suitable as the vocabulary unit to teach, learn and measure at early stages of vocabulary development, while *lemmas* are appropriate at late beginner and intermediate level and *word families* are more suitable among advanced EFL learners and native speakers.

In the current study, the participants are EFL learners who are expected to have a low but non-beginner level of vocabulary knowledge. Thus, it would be more suitable to use lemmas as the unit for teaching and counting words. In fact, the current study used the Vocabulary Size Test (Nation & Beglar, 2007) to measure the vocabulary size of the participants, which is a ready-made professional instrument that claims to measure *families*. For the teaching and testing of the words that were used in the experiment, however, the current study works with *lemmas*, although actually in the exercises and tests, inflected or derived forms of the chosen words were never used. The target words all appear only in base form, which for verbs is either infinitive, imperative, or present

[4] Cobb, T. Familizer + Lemmatizer v.2 [computer program]. Accessed 15 April 2019 at https://www.lextutor.ca/familizer/ .

simple other than third person singular. Hence in fact it could be equally said that the current study concerns only the teaching and learning of base form word *types*.

2.4 The nature of vocabulary knowledge

Vocabulary knowledge is a rather complex concept. The general consensus, however, is that vocabulary learning is not an all-or-nothing process, but entails many types or degrees of vocabulary knowledge and requires multiple exposures to obtain anything approaching full knowledge (Anderson, & Herman, 1987; Henriksen, 1999; Nagy, & Scott, 2000; Nation, 2001; Hirsch, 2003; Hunt & Beglar, 2005; Joe, 2010; Schmitt, 1998, 2000, 2010). Because there are different degrees to vocabulary knowledge, it is commonly described via a number of distinctions.

The first distinction is between vocabulary *breadth* and *depth*. *Breadth* refers to the number of vocabulary items known, at least minimally, while *depth* refers to the quality or how much a learner knows of a word. This distinction, however, could be criticized as both the terms *breadth* and *depth* may carry various meanings (Milton, 2009). For example, *breadth* must be accompanied by some degree of depth and so may refer only to a learner’s ability to recognise a word’s form, or to a learner’s ability to link a word’s form with its meaning, or to recognise a word and link it to a translation in the first language (L1) etc. (Milton, 2013). It is even more difficult to narrow down what is involved in *depth* of vocabulary knowledge because it may refer to knowledge of word associations (e.g., Meara, 1983; Fitzpatrick, 2006), grammatical functions (e.g., DeKeyser, 2005), collocations (e.g., Barfield, 2005; Barfield & Gyllstad, 2009), synonyms (e.g., Qian, 2008), multiple senses (e.g., Schmitt, 1998), and many other kinds of knowledge, including its word family.

The second distinction is between *receptive* and *productive* vocabulary knowledge. Receptive vocabulary knowledge refers to the ability to understand a word in reading or listening while productive vocabulary knowledge refers to the ability to use a word in writing or speaking. This distinction does not necessarily mean that receptive vocabulary knowledge and productive vocabulary knowledge are distinct, but rather occupy different points on a continuum (Melka, 1997; Henriksen, 1999; Read, 2000; Schmitt, 2000, 2010). The idea of a scale arises from the fact that productive vocabulary size usually lags behind receptive vocabulary size (Laufer & Goldstein, 2004) as words tend to be initially known receptively, where less depth of knowledge is required, then over time become known productively, where a fuller knowledge of different aspects of the word, such as its grammar and collocation, is required (Schmitt, 2010). Furthermore, some studies suggest that productive vocabulary knowledge declines faster than receptive vocabulary knowledge after learning has ceased (Schneider, Healy, & Bourne, 2002). It should be noted however that in practice it is difficult to identify the threshold between *receptive* and *productive* vocabulary knowledge (Meara, 1997; Henriksen, 1999) and the exact *depth* of vocabulary knowledge which is required to facilitate the move from receptive to productive vocabulary knowledge (Schmitt, 2010).

The third distinction, and perhaps the one that gives a more articulated picture of vocabulary knowledge, was first expressed by Richards (1976), and further elaborated on by Nation (2001). In this distinction, vocabulary *depth* knowledge which was sketched above is divided into knowledge of *form*, knowledge of *meaning*, and knowledge of *use*. Each of these divisions are further subdivided into three aspects and each has a receptive and productive mode (Table 2.2).

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognisable in this word?
		P	What words parts are needed to express meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this word make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of word occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use	R	Where, when and how often would we meet this word?
		P	Where, when and how often can we use this word?
R: Receptive vocabulary knowledge P: Productive vocabulary knowledge			

Table 2.2 What is involved in knowing a word (Nation, 2001, p. 27)

According to Nation (2001), Knowledge of *form* includes knowledge of the phonological form, knowledge of the written form and knowledge of word parts. Knowledge of phonological form involves knowing the sound of a word and its pronunciation. Knowledge of the written form simply refers to knowing what a word looks like and/or how it is written and spelled. Knowledge of word parts involves knowledge of affixation as in recognizing and/or using parts of a word that change the meaning of a word. For example, knowledge of word parts involves knowing that the meaning of many words can be negated by adding the prefix *dis*-, as in *advantage* and *disadvantage*.

Knowledge of *meaning* involves knowledge of the connection of form and meaning, knowledge of concepts and referents, and knowledge of associations. Knowledge of form and meaning refers to the ability to link between a form of a word and its meaning, which perhaps for EFL learners involves linking the form of a word to its translation in the learners' mother tongue. However, the same word in different languages might carry different concepts and associations (Milton, 2009). In other words, learners require more than a simple knowledge of how to link between form and meaning, because using words appropriately involves knowing the concepts and associations that words carry with them.

Knowledge of *use* involves knowledge of grammatical functions, collocations, and constraints on use. The knowledge of grammatical functions involves knowing the part of speech of a word and how to use a word within a sentence or sequence. For example, most adjectives in English occur in a noun phrase (e.g., a *big* garden, an *old* house) and after copular verbs (e.g., the garden is *big*, the house is *old*). However, there are some adjectives that can only occur in the noun phrase (e.g., the *main* reason, a *true* friend) and not after copular verbs (e.g., the reason is main, a friend is true). There is no simple rule to explain why some adjectives can occur after copular verbs, while other adjectives cannot. As such, English learners have to learn each one of these adjectives.

Knowledge of collocations refers to the knowledge of the tendency of some words to co-occur with each other. Some words tend to occur very frequently with other words such as the English adjective *heavy* which collocates with the nouns *burden*, *drinker*, and *cold.* Some verbs also are restricted in their collocability, for example, the English verb *commit* collocates with the noun *murder* and could only be replaced with another less frequent verb such as *perpetrate.* Moreover, other words such as *good* or *buy* have less restriction and could collocate freely with a wide range of other words. The final aspect of word knowledge is constraints on use. Nation (2001, p. 27) did not fully elaborate on how to distinguish constraints on use but generally described them in the questions “Where, when and how often we would meet this word?” and “Where, when and how often we can use this word?”. Furthermore, he generally identifies *register* and *frequency* as the most important constrains. Some words, for example, are more appropriate in a formal context than other words such as in business letters or academic writing.

In the present study it is necessary to make some choices from all the above distinctions, as it is impossible to investigate the teaching/learning of all of them in one study. The study will deal with both receptive and productive knowledge of vocabulary, but it will limit attention to just a few aspects of depth of vocabulary knowledge. These aspects will be essentially just the form-meaning relationship for the written form, whose length will be controlled. The study will not be concerned with spoken forms, all word parts, concepts, associations, grammatical patterns, collocation, or constraints on use, where frequency will in fact be controlled. The form-meaning relationship will however be learned, practised, and tested in two ways (recognition and recall) which will be described in a later section (Section 2.10.5.1).

2.5 Instructed second and foreign language vocabulary learning

Vocabulary learning is a rather complex concept as it involves a wide range of features (de Groot, 2006, Milton, 2009) and it is unclear what is the best means of learning vocabulary (Laufer &

Roitblat, 2011). In general, however, vocabulary learning can be classified into two learning modes: *incidental* and *intentional* vocabulary learning (Nation, 2001, Hulstijn, 2003).

There are a few terminological considerations however that need to be addressed before defining the terms *incidental* and *intentional* learning. In the applied linguistics literature, the terms *incidental* and *intentional* learning are occasionally used interchangeably with the terms *implicit* and *explicit* learning, respectively (Hulstijn, 2013). It is generally recommended however that a distinction must be made between incidental and implicit learning, as well as, between intentional and explicit learning (Schmidt, 1994, Hulstijn, 2003).

In psychology, implicit and explicit learning are largely distinguished based on the absence or presence of consciousness when learning takes place, while incidental and intentional learning are characterised based on whether there is an intention or not for learning some kind of information, in our case vocabulary (Rieder, 2003). Generally, implicit learning can be defined as the unconscious learning process that takes place without a learner's awareness through mere exposure to input (Rieder, 2003; Hulstijn, 2005), while explicit learning occurs "when the learner has online awareness, formulating and testing conscious hypotheses in the course of learning" (Ellis, 1994, p. 38). Incidental learning can be defined as the unintentional learning process taking place with or without the learner's awareness (Rieder, 2003), while intentional learning refers to deliberate and conscious processes that have learning of some target information as the goal (Leow & Zamora, 2017). As Milton (2009) points out, the distinction between implicit and explicit learning is usually ignored in language learning terminology, while the terms incidental and implicit learning are often used synonymously in language teaching.

In line with the above discussion, it is possible to view incidental vocabulary learning as consisting of implicit and/or explicit learning processes and view intentional vocabulary learning as being composed of only explicit processes (Rieder, 2003). Thus, *Incidental vocabulary learning* can perhaps be described as any undeliberate learning that takes place by ‘picking up’ new words and committing them in memory through exposure to written and spoken texts, such as reading for pleasure or listening to the radio (Hulstijn, 2013). The aim or intention of the learner in that case is typically to understand and enjoy the message of the input. When unknown words inevitably arise, these words hinder that intention and lead to the learner doing some conscious work by guessing or using a dictionary so as to be able to continue with the intended task. The vocabulary learning is therefore not intended, but is still done at a conscious, explicit level. In incidental vocabulary teaching, the aim is to attract an L2 learner’s attention to the meaning of new words (Khezrlou, Ellis, and Sadeghi, 2017), consciously or not, through multiple exposures to these words in a wide range of tasks and contexts (Nation, 2001) and without forewarning the learner of any upcoming vocabulary test (Hulstijn & Laufer, 2001).

On the other hand, *Intentional vocabulary learning* can be described as a deliberate process of finding out new information about words and expressions and storing it in memory (Hulstijn, 2013), where that is the primary aim of the task. This process may be done by the learners themselves, where for example they use conscious vocabulary learning strategies such as keeping notes of new words and deliberately going over them using repetition or association strategies so as to learn them. Alternatively, intentional learning may occur through deliberate instruction by the teacher, using vocabulary learning activities (Fernandez & Schmitt, 2017), such as multiple-choice, translation, and gap-filling tasks. In intentional vocabulary teaching, the attention of an L2 learner is directed to form and meaning as the main aim of the task (Sonbul & Schmitt, 2010) and the learner may be aware of an upcoming vocabulary test (Hulstijn & Laufer, 2001).

It is a commonly held belief that incidental vocabulary learning leads to higher gains in L2 vocabulary knowledge than intentional learning and instruction (e.g., Nagy, Herman, and Anderson, 1985; Nagy and Herman, 1987; Harris & Snow, 2004). Some researchers have claimed that L2 learners would acquire large amount of words without any explicit learning or teaching by simply being exposed to the L2 language. However, recent research suggests that the misconception about the benefit of intentional vocabulary learning, as opposed to incidental vocabulary learning, is “based on an ill-informed understanding of the terms incidental and intentional learning” (Hulstijn, 2013, p. 4). Typically, gains in vocabulary knowledge are usually modest and slow when learning is incidental and implicit (Nation, 2001; Read, 2004; File & Adams, 2010; Khezrlou, Ellis, & Sadeghi, 2017), whereas vocabulary gains are often rapid and large when learning is intentional and explicit (Nation, 2001; Lehmann, 2007; Schmitt, 2008).

According to Groot (2000, p. 59), learning new words only through incidental exposure to L2 input is unsuitable for foreign language learners for several reasons. First, low frequency words would occur rarely in small authentic texts. Therefore, there is simply not enough repetition to allow for the various features of words to be learned from varied contexts, which negatively affects the incremental learning process of these words. For example, it is estimated that language learners would have to read over 200,000 words of texts to facilitate incidental learning of 108 words and read a text of around eight million words to develop a vocabulary size of 2,000 words (Nation & Wang, 1999). However, it would be unrealistic to expect learners to read such amounts of text in most educational settings, especially instructed foreign language contexts such as Saudi Arabia where learners may do little or no reading and listening to English outside the classroom.

Second, incidental and implicit learning of words must be done through guessing/inferencing, since use of a dictionary inevitably involves conscious attention and must be explicit. However, inferring the meaning of unknown words in authentic texts often relies for success on wider contexts not immediate contexts. In most instructed L2 learning settings, however, learners are usually exposed only to small reading passages, which may not provide the learners with many cues to the meaning of unknown words. Indeed, studies have reported that learners usually fail in guessing the meaning of unknown words from context with a success rate ranging from 9.5% to 38.1% (Nassaji, 2003), and “the more often the word was correctly guessed, the less often it was remembered" (Mondria & Boer, 1991, p. 262). Furthermore, unless texts are very carefully chosen, in a way that EFL learners would not be able to do for themselves, there might well be too many unknown words in the text which reduce the chances of successful contextual deduction and, consequently, incidental acquisition. For this reason, L2 learners need to already have a large vocabulary size to learn new words from incidental exposure to authentic texts (Horst, Cobb, & Meara, 1998), which probably suggests that incidental vocabulary learning is more likely to be beneficial for highly advanced language learners (Bowne, Yoshikawa, & Snow, 2017).

As discussed in Section 2.3, L2 learners need to develop a large lexicon to achieve adequate comprehension of authentic written and spoken texts. In first language acquisition, most words are indeed learned incrementally through incidental and implicit exposure to a wide range of spoken and written language (De Bot, Paribakht, & Wesche, 1997). In foreign language acquisition, however, language learners cannot duplicate the natural process of L1 vocabulary acquisition as multiple exposures to new words are substantially less obtainable and incidental learning may only occur for a limited number of very frequent words. There is simply not enough time in most L2 educational settings for learners to develop a large vocabulary size through incidental exposure to authentic L2 contexts.

Overall, incidental vocabulary learning should not therefore be considered as a primary source of foreign language vocabulary learning (Schmitt, 2008). Incidental learning through reading does seem to promote vocabulary learning, albeit pickup rate is slow, low, and not effective for developing productive knowledge (Brown, Waring, & Donkaewbua, 2008). It is generally recommended therefore that incidental vocabulary learning should be enhanced by integrating an explicit and intentional learning element into reading activities, (e.g., post-reading multiple-choice task focusing on the words to be learnt) (Hill & Laufer, 2003). On its own, intentional vocabulary learning indeed usually leads to rapid and large gains of vocabulary knowledge, "with a better chance of retention and of reaching productive levels of mastery" (Schmitt, 2008, p. 341), especially with low proficiency learners (Nation, 2001; Lehmann, 2007). As Milton (2009, p. 2) succinctly puts it, “successful learners acquire large volumes of vocabulary from the words explicitly taught in the classroom and supplement their learning by targeting vocabulary in activities”.

The advantage of more explicit learning and/or instruction can be explained in terms of Schmidt’s (2001; 2010) *noticing hypothesis*, which suggests that when learners pay attention to input, they process this input and convert it to intake for learning. The role of ‘noticing’ and attention in vocabulary learning has recently been likened to a pedagogical approach from the area of grammar research known as *form-focused instruction* (Long, 1991; Ellis, 2001), which would be *word-focused instruction* in vocabulary learning (Laufer, 2005). Form-focused instruction can be divided into two main pedagogical approaches, namely *Focus on Form* (FonF) and *Focus on Forms* (FonFs) (Long, 1991, 2009).

In the case of vocabulary learning or instruction, FonF entails both explicit and implicit vocabulary learning which is incidental to the main aim of the task being performed (Ellis, 2001). FonF can be achieved by attracting a learner’s attention to words during communicative tasks (Laufer and Rozovski-Roitblat, 2011), such as looking-up the meaning of an unknown word in the dictionary while reading a text. FonFs, on the other hand, occurs where the main aim of the task is vocabulary oriented. Here there is intentional and explicit “vocabulary practice of discrete lexical items in a noncommunicative, nonauthentic environment” (Laufer, 2006 p. 150). For example, learning new words by the FonFs method can be achieved through repeating word lists, or completing multiple-choice exercises. According to Ellis (2001, p. 14), the underlying assumption in FonF is that learners view themselves as users of the language and pay attention to certain linguistic features only incidentally as they occur in tasks that primarily focus on message communication, whereas in FonFs the learners view themselves as L2 learners rather than users and treat the L2 language as an “object” that they study and practice.

In general, the relative effectiveness of the different types of form-focused instruction (e.g., FonF, FonFs), has mostly been examined in terms of the degree of explicitness of instructional intervention (Graaff & Housen, 2009). Table 2.3 summarises the key features that have been used to distinguish implicit and explicit form-focused instruction.

Implicit FFI	Explicit FFI
• attracts attention to language form.	• directs attention to language form.
• language serves primarily as a tool for communication.	• language serves as an object of study.
• delivered spontaneously and incidentally (e.g., in an otherwise communication-oriented activity). FonF	• predetermined and planned (e.g., as the main focus and goal of a teaching activity). FonFs
• unobtrusive (minimal interruption of communication of meaning).	• obtrusive (interruption of communication of meaning).
• presents target forms in context.	• presents target forms in isolation.
• no rule explanation or directions to attend to forms to discover rules; no use of metalanguage.	• use of rule explanation or directions to attend to forms to discover rules; use of metalinguistic terminology.
• encourages free use of target form.	• involves controlled practice of target form.

Table 2.3 Features of implicit and explicit forms of form-focused instruction (Graaff & Housen, 2009, p. 737)

The majority of empirical research on form-focused instruction has been conducted in the area of grammar instruction. Only recently have more studies investigated form-focused instruction in relation to different types of vocabulary instruction (e.g., Laufer, 2005, 2006; Peters, 2006; Laufer

& Girsai, 2008; Laufer and Rozovski-Roitblat, 2011, 2014). Laufer (2005), surveyed a number of studies on vocabulary learning through word-focused tasks, with or without reading, and identified three types of word-focused instruction (see Figure 2.1). According to Laufer (2005), any type of word-focused instruction, which she terms task embedded, task related, and task unrelated FFI, is better than no word-focused instruction at all. Laufer (2017) further asserts that a word-focused instruction approach is indispensable in vocabulary instruction and plays a vital role in the development of the learner's lexical knowledge (Laufer, 2017).

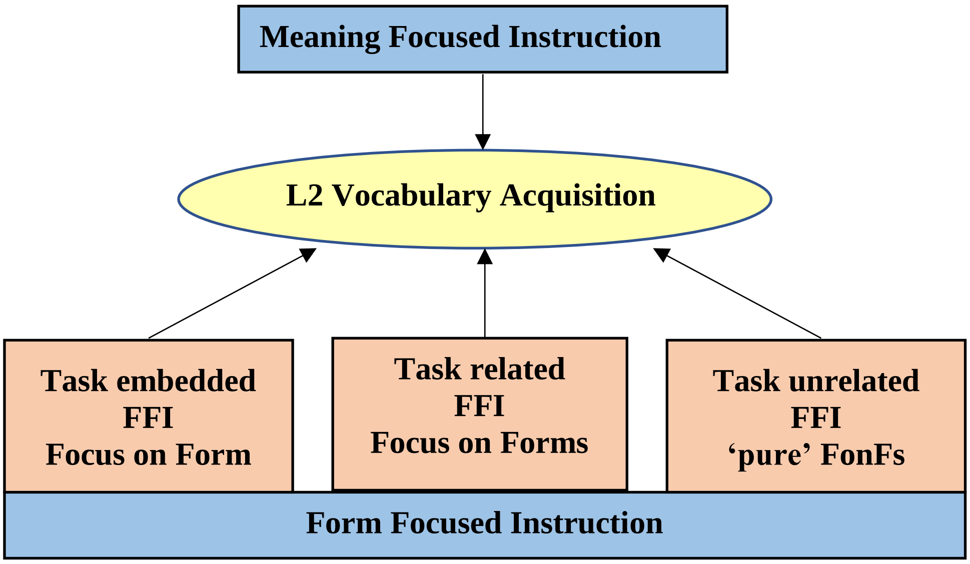


Figure 2.1 L2 components of vocabulary instruction (Laufer, 2005, p. 145)

Non-communicative and decontextualized word-focused activities have lost their popularity over the past few decades due to the shift towards communicative or task-based vocabulary learning activities (Peters, 2014). However, research has shown that explicit FonFs vocabulary learning activities lead to higher vocabulary gains than FonF activities. For example, Laufer (2006) investigated the effect of FonF and FonFs on incidental and intentional vocabulary learning among L1 Arabic or Hebrew secondary school learners of English. In the incidental learning session, FonF was operationalized through asking the student to read a text and use a dictionary as needed. In the intentional learning session, FonFs was operationalized by explicitly teaching the students a list of 12 words (i.e., L1 translations and English forms) and practicing these words using word-focused tasks (e.g., multiple-choice, and gapped sentences exercises). Learning was measured by a productive recall test (i.e., L1-L2 translation test) and a receptive recall test (i.e., L2-L1 translation test). Results indicated that intentional learning led to higher vocabulary retention than incidental learning, and that the students in the FonFs condition outperformed the students in FonF condition. Laufer (2006, p. 162) argued that, in any learning context, "the nature of lexical competence makes FonFs indispensable to vocabulary instruction". Similar findings were reported in Keating (2008) and Laufer and Rozovski-Roitblat (2011).

In the present study, where the focus is on the spacing of the repeated vocabulary tasks rather than on comparing different kinds of vocabulary learning/use tasks, a decision had to be made as to what the actual learning task would be. As will emerge later, spaced/massed classroom research always seems to focus on learning or instruction that is fully explicit and intentional. Hence for that reason, as well as the proven superiority of that kind of vocabulary learning just mentioned, I have chosen fully FonFs tasks for the present study. That is to say that the students will be practising new vocabulary information very much as described on the right-hand side of Table 2.3 and Figure 2.1.

2.6 Factors potentially affecting vocabulary learnability that are relevant to the present study

Research on L2 vocabulary acquisition suggests that many factors may influence the ease or difficulty of instructed vocabulary learning. These factors include ones related to the learner (e.g., motivation, proficiency level, strategic competence, age, prior knowledge of target words etc.), ones related to the classroom context (e.g., the teacher's communication style, whether tasks are done by students individually or in groups, etc.), ones related to the chosen task (e.g., task’s incidental or intentional nature, inherent ease or interest as an activity, and whether it involves recall or recognition of information), and ones related to the words and lexical information about them that are to be learned.

Intrinsic features of the word itself which may influence vocabulary learnability include: part of speech (e.g., Horst & Meara, 1999), cognateness (e.g., Nation 2006; Tonzar, Lotto, & Job, 2009) and hence potential language transfer (e.g., Koda, 1997; de Groot, 2006), L1 lexical translation equivalence (e.g., Al-Masrai & Milton, 2015b), imageability of meaning (e.g., Ellis & Beaton, 1993), word length (e.g., Willis & Ohashi, 2012), and frequency (e.g., Richards & Malvern, 2007). Other such factors include inflexional complexity, derivational complexity, contextual restrictions, and similarity of lexical forms within L2 (Laufer, 1997). In addition, the variable of repetition (e.g., Rott, 1999; Webb, 2007) is clearly crucial in the current study.

The following subsections briefly review a few of these learnability factors that may have a role in the ease or difficulty of vocabulary learning in the current study.

2.6.1 Word class

A number of studies investigated the effect of word class as a factor contributing to vocabulary learning difficulty. These studies do not seem to unanimously reach the same conclusion. For instance, some studies suggest that noun learning is less demanding than verb learning, whereas other studies suggest that verbs are easier to learn than nouns. According to Maguire, Hirsh-Pasek, and Golinkoff (2006, p. 367) “mapping from action or mental state to word is considerably more challenging than mapping from object to word”. More recently, McNamara, Crossley, and Roscoe (2013) suggest that nouns possess many properties (i.e., imageability, concreteness, specificity, meaningfulness) which differ from verbs which are ambiguous and abstract. As such, noun learning can be achieved through multiple exposures alone, while verb learning requires multiple exposures and contextual diversity.

This claim seems to be in line with a few studies. For example, Ellis and Beaton (1993) cite Rodgers (1969), which examined EFL vocabulary list-learning and found nouns to be the easiest part of speech to learn, followed by adjectives then verbs and adverbs as the most difficult to learn. Similarly, Laufer (1997) and Horst and Meara (1999) explored the role of word class in the ease or difficulty of L2 vocabulary learning. These two studies reported that nouns were easier to learn, followed by verbs and adjectives, then adverbs as the most difficult.

On the other hand, other studies seem to suggest that noun learning is not the least demanding part of speech of vocabulary learning. For example, Schwanenflugel, Stahl, and Mcfalls (1997) found that verbs, adjectives, and adverbs were easier to learn than nouns. Al Fotais (2012), suggested that adjectives were easier to learn followed by nouns then verbs. Al-Masrai and Milton (2015a) examined word difficulty and learning among Saudi EFL learners. The study reported that word class did not have a clear effect on vocabulary learnability. In light of these studies, it could be possible to assume that word class may not have a clear effect on the difficulty or ease of vocabulary learning, which is in line with Laufer (1997). In any event, the lack of clarity in the research just mentioned encouraged me to include word class as a variable in my study, since it is clearly a variable that needs further attention in vocabulary learning research.

2.6.2 Word length

It is generally assumed that longer words are more difficult to learn than shorter words because there is “more to remember in long words than in short words” (Nation & Webb, 2011, p. 315). However, it should be noted that the literature on word length and its relationship to the ease or difficulty of vocabulary learning does not yield conclusive findings (e.g., Laufer, 1990, 1997;

Milton, 2009; Al Fotais, 2012; Al-Masrai & Milton, 2015a). Several studies did not find a clear relationship between word length and vocabulary learning. For example, Milton and Daller (2007) and Milton (2009) suggested that word length in syllables did not show a significant effect on the learning of French vocabulary among British learners. Similarly, Al Fotais (2012) suggested that there was no clear effect between word length in letters and vocabulary learnability among Saudi EFL learners. More recently, Koirala (2015) reported that Spanish and Portuguese EFL learners rated words of varying numbers of letters and syllables as almost equally easy to learn.

On the other hand, several studies reported an effect of word length on vocabulary learning. For example, Alsaif and Milton (2012) examined the effect of frequency, cognateness, and word length in syllables on vocabulary learning among Saudi EFL learners in public schools. They reported that all three factors combined have an impact on vocabulary learning and accounted for 63.8% of variance in the participants' overall scores. Interestingly, the study found that word length had the largest impact on vocabulary learning and accounted for around 36% of the aforementioned variance. This finding contradicts Milton and Daller (2007) in which word length in syllables did not show a significant effect on the learning of French vocabulary by British learners. According to Alsaif and Milton (2012), this finding could be explained in terms of the degree of similarity between the participants' L1 and L2 in these studies. For example, in Alsaif & Milton (2012) the words were tested in base forms only (i.e., types), while in Milton and Daller (2007) the words were tested in base forms and inflected forms (i.e., lemmas). Therefore, due to the similarity of affixes in English and French it could be possible for participants in Milton (2009, p. 41) "to reduce an unknown long word to shorter component parts that are either known or are guessable".

Willis & Ohashi (2012) is a partial replication of Milton and Daller (2007). The study examined the effects of word frequency, cognateness and word length in letters, syllables, and phonemes on L2 English vocabulary learnability. The participants of the study were L1 Japanese university students. The study used the first seven levels of the Vocabulary Size Test (VST) (Nation &

Beglar, 2007) to measure the participants' receptive recognition vocabulary. Results suggest that frequency, cognateness, and length in phonemes best predicts L2 vocabulary learnability for Japanese ESL learners. Particularly, vocabulary learnability seems to be largely affected by cognateness, followed by frequency and then by word length in phonemes.

According to Laufer (1997), the morphological transparency of longer words makes it difficult to determine the effect of length in many studies. For example, a long word such as *mismanagement* consists of familiar morphemes which could be easily decoded and learned by the learner. Furthermore, the incongruent findings regarding the effect of word length on vocabulary learnability could be attributed to how different studies operationalised word length; in other words, different length measures might have produced different results (Milton, 2009). For example, word length has been operationalized as the number of letters (e.g., Al Fotais, 2012; Koirala, 2015), the number of syllables, either written or spoken (e.g., Milton & Daller, 2007; Milton, 2009; Alsaif & Milton, 2012; Al-Masrai & Milton, 2015a), or the number of phonemes (e.g., Willis & Ohashi, 2012). Consequently, it would reasonable to assume that word length is a factor that should be considered when designing studies that investigate vocabulary learning.

In the present study, however, word length was not treated as a variable, as it would lead to the danger of the study becoming unmanageable due to too many independent variables. Rather, the current study took steps to eliminate its effect by keeping it constant across all the target words that were used.

2.6.3 L1 lexical translation equivalence

This refers to how far a word has a single word translation equivalent in the learner's L1 which possesses exactly the same meaning. This may influence learnability of FL vocabulary regardless of how words are presented to learners by a teacher or textbook, or how far translation is used by the learner as a learning strategy. However, it is usually seen as likely to be more influential if translation is heavily and explicitly used in the teaching/learning process.

The use of translation in L2 learning is in fact often criticised by language teachers as it is considered a remnant of the Grammar translation method and not supported by more preferable and recent language teaching methods, such as the Communicative language teaching approach (Hummel, 2010; Hall & Cook, 2012). As such, translation is generally viewed as being an unhelpful method of L2 learning and impeding the use of the L2 inside the classroom. According to Nation (2013), however, most other methods of conveying meaning inside the classroom suffer from the same criticisms levelled against translation. For example, L2 words do not have exact L1 equivalents but in the same way L2 words do not have exact L2 definition equivalents either. Furthermore, visual aids and demonstrations may negatively impact the use of the L2 inside the classroom in a similar manner to when the meaning is communicated through L1 translation. Indeed, translation can actually be seen as an effective pedagogical tool in L2 vocabulary learning as it is quick, simple and caters for large classes (Nation, 2013). Furthermore, learning vocabulary through translation tasks could be more effective than other methods, such L2 definitions, due to the fact that an L1 translation is “fully familiar to the learner and consists of only one word” (Laufer & Shmueli, 1997, p. 103) whereas, an L2 definition may consist of long phrases and unfamiliar words that distract the learner’s attention from the target words.

In any event, in Saudi Arabia, as in many EFL countries, translation remains considerably used in vocabulary presentation and practice in the classroom and is widely used by learners themselves in their own dictionary lookup and learning of new words. Studies on vocabulary learning strategies indicate that learning L2 words along their L1 translations was the most frequent strategy among Saudi EFL learners (Al-Akloby, 2001; AlQahtani, 2005; Alhatmi, 2012). Psycholinguistic research also suggests that L2 learners with low L2 proficiency spontaneously tend to directly link new L2 words to their L1 equivalents, especially at initial stages of vocabulary learning (Jiang, 2000, 2004). Hence, the learnability impact of L1-L2 vocabulary differences is addressed here, which may have a greater effect the more translation is used, through the process known as 'transfer'.

In the early years of the examination of the effect of the L1 on L2 learning, the general view was that it was the differences between L1 and L2 structures that accounted for crosslinguistic transfer and so affected learnability (Stockwell, Bowen, & Martin, 1965). However, this view was later opposed in the 1970s and 1980s and, instead, the similarities between the L1 and L2 were suggested to account for L1 transfer (Kellerman, 1983; Ringbom, 1987). The current view of crosslinguistic transfer, however, is that it can be a product of both similarities and differences between the L1 and the L2 (Kellerman, 1995). Accordingly, crosslinguistic transfer can be defined as “the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired” (Odlin, 1989, p. 27) and “can be observed in all linguistic subsystems, or at all levels of language use” (Piasecka, 2006, p. 246).

In more detail, research investigating the effect of crosslinguistic transfer on language learning has identified several types of transfer: especially *positive transfer* and *negative transfer*. Other types of crosslinguistic transfer have also been suggested by James (1994, p. 183), who distinguished between *primary transfer*, which refers to “the spontaneous, untaught strategy of each individual learner” and *secondary transfer*, which refers to transfer as a product of the individual’s “legacy from the community in a language contact situation” (James, 1994, p. 183). For present purposes however, we need not pursue those separately. Positive transfer refers to when learners transfer equivalent lexical structures from the L1 to the L2 (Laufer, 1992; Ringbom & Jarvis, 2009), which would make an L2 word easier to learn, while negative transfer usually refers to transferring non-equivalent lexical structures from the L1 to the L2 (Piasecka, 2006; Gor & Vatz, 2009), which makes it harder to learn the correct L2 forms and meanings.

For example, orthographical and phonological similarities between L1 and L2 have been reported to promote learning new vocabulary (Ellis & Beaton, 1993; de Groot, 2006); on the other hand, differences in those respects may make learning more difficult. In the case of Arabic and English there is no similarity of writing system to assist learning words, but that affects the learnability of all English words more or less equally. There are also many phonological differences between the two languages, but these would not affect all English words in the same way. For instance, due to the incongruent phonological systems of the L1 and L2, a common error of Arab EFL learners can be observed in pronouncing the letter [p] as a voiced bilabial stop /b/ instead of a voiceless bilabial stop /p/ (Aljasser, Jackson, Vitevitch, & Sereno, 2018). That makes English words containing /p/ harder to learn than those without /p/.

Where vocabulary is concerned, however, the most often discussed areas of similarity and difference between languages, leading to positive or negative transfer and so differential word learnability, fall into two areas: cognates/false friends, and degree of one to one translation equivalence of words. The first need not to be pursued, since it focuses on instances of words with a similar form (especially sound shape) in the L1 and L2 and there is quite low incidence of this between English and any variety of Arabic, so issues of how positive or negative the effects of this are on learning an English word rarely arise. The second however occurs everywhere in any pair of languages and is the main source of differential learnability considered in the current subsection.

Languages do not share the same conceptual systems and hence it is quite common (a) for words in two languages to constitute translation equivalents of each other only in some respects, not in all senses and contexts, or indeed (b) for words to exist in one language for which there is no single word equivalent in the other at all. In general, low proficiency EFL learners are expected to find it easier to learn new L2 words that have an existing L1 translation equivalent than new L2 words that require the construction of new conceptual knowledge (Milton, 2009). As such, translation equivalency at the word level could have an impact on the ease or difficulty of vocabulary learning.

For example, the English word ‘couple’ is translated as ‘*zawjan* زوجان’ in Modern Standard Arabic, but the two words are not fully equivalent because a ‘couple’ in English refers to two persons who are married, engaged, or romantically paired, while ‘*zawjan* زوجان’ in Arabic can only mean two persons who are married. Therefore, it is necessary that an L2 learner “readjusts the semantic knowledge of the word that s/he possesses to that of the native speaker” (Laufer & Girsai, 2008, p. 699) in order to use the new L2 words correctly. By contrast, a word like ‘breeze’ has a straightforward equivalent in Arabic ‘*nassem* نسيم’ and so would be regarded as relatively more learnable than *couple*.

According to Ringbom (2007), the extent to which translation equivalency may influence L2 word learning is dependent on a learner’s perceived proximity (i.e., resemblance) between the L1 and L2. This subjective notion of language proximity is usually referred to by researchers as *psychotypology* (Kellerman, 1983). Ringbom (1978) found that Finnish and Swedish bilinguals learning English as their third language generated most errors due to transfer from Swedish grammatical rules to English but not Finnish. As such, the effect of transfer was present based on only one of the two L1s. This finding was explained by Ringbom (1978) as a result of the learners’ different perceived psychotypological distances of English from the two L1s, which led the learners to transfer from Swedish, which is historically as well as typologically related to English, while suppressing transfer from Finnish, which is not historically or typologically related to English. Jordens and Kellerman (1981) found similar results in a study of acceptability of Dutch idiomatic expressions by two groups of Dutch native speakers, one group learning English and the other group learning German. Results suggested that learners tended to accept German idioms, which are more typologically related to Dutch, while reject English idioms, which are not typologically related to Dutch. Such results, therefore, may indicate that L1 transfer may at times rely on the psychotypological perception by learners. In the current study, however, since all the participants speak only varieties of Arabic and not any other languages such as French that are typologically closer to English, this will not be a variable factor. All participants would be expected to view English as typologically distant from Arabic and so be relatively disinclined to transfer from L1.

In a recent study, however, Al-Masrai and Milton (2015a) investigated word difficulty and learning among final-year Saudi EFL high school students. The study examined L1 translation equivalency as a difficulty factor along with amount of vocabulary recycling in textbooks, word length (i.e., the number of syllables), and part of speech. The participants were given a receptive form recognition test (i.e., yes/no test), a receptive meaning recall test (i.e., English-Arabic translation test), and a productive meaning recall test (i.e., Arabic-English translation test). A regression analysis indicated that word class and word length did not have a clear impact on vocabulary learning, while textbook recycling accounted for 60% of the variance and word translation equivalence accounted for 23% of the variance in the overall model of learning. This finding suggests therefore that, regardless of the impact on transfer of psychotypology as suggested in the previous paragraph, L2 English words that have straightforward L1 translation equivalents were more likely to be learned than words with no L1 translation equivalents. In another study, Al-Masrai and Milton (2015b) further examined the relationship between L1 word translation equivalence and L2 vocabulary learning among final-year Saudi EFL high school students. The participants were given a receptive form recognition test (i.e., yes/no test) and a productive meaning recall test (i.e., Arabic-English translation test). Results indicated again that there was a significantly greater uptake of L2 words with L1 translation equivalence than L2 words which do not have translation equivalents.

Since L1 translation equivalency could have an impact on vocabulary learnability among Saudi EFL learners, it might be important for vocabulary studies to “control for it as much as possible in a research design and to consider its effects in the interpretation of study results” (Schmitt, 2010,

p. 75). Hence in the present study this factor was controlled by choosing for the experiment only English words which had straightforward Arabic equivalents.

2.6.4 Frequency of occurrence

A common observation from research into word frequency suggests that language learners are more likely to acquire words that occur more frequently in the language (Schmitt, Schmitt, and Clapham, 2001, Schmitt, 2010). Furthermore, word frequency not only seems to affect how easy a word is to learn but also how prone it is to attrition (Milton, 2009). *Word frequency* refers to the rate of occurrence of a word in written or spoken text. However, it should be noted that, word frequency may vary depending on how a researcher decides to count the rate of occurrence of a word in a text, including what sort of unit he/she chooses to count as a word (see Section 2.3). Also, crucially word frequency depends on what text is chosen, especially whether it is text that represents the actual input learners have received (e.g., their textbook or what they read), or text that represents language use by some large group of people such as native speakers (e.g., the British National Corpus).

In a comprehensive study, Reynolds and Wible (2014) examined six studies that investigated the effect of word frequency on vocabulary acquisition through incidental reading, namely, Horst, Cobb, and Meara (1998), Zahar, Cobb, and Spada (2001), Waring and Takaki (2003), Tekmen and Daloğlu (2006), Brown, Waring, and Donkaewbua (2008), and Pellicer-Sánchez and Schmitt (2010). The results indicated that different researchers operationalised word frequency differently and that, in each study, different methods of word counting were used for counting different words in the same text. As such, Reynolds and Wible (2014, p. 858) concluded that despite increased interest in word frequency as a research area, the operationalization of word frequency has "remained implicit and, in some cases, inconsistent".

It is well established that the number of English 'words' in a text may vary from several thousand words to a few hundred words depending on what is defined as a 'word'. As previously discussed in Section 2.3, different studies adopted different definitions of the term word, which not only might have exaggerated early estimates of native speaker’s vocabulary size (Nation, 1993), but also affect studies of frequency as a factor in word learnability. Traditionally, word frequency has been used by syllabus and textbook writers as a criterion to decide which words are more important to learn than others, especially at early stages of vocabulary acquisition. Words like *the*, *and*, *to*, *it* and *for* occur very frequently so that almost any written or spoken text will include them. Other words like *pleura*, *pizzicato*, *neaten,* and *oligotrophic* occur very infrequently that almost any written and spoken text will exclude them. As such, especially at early stages of vocabulary acquisition, language learners will most likely encounter a lot of very frequent words, which makes frequency the most important criterion in choosing the vocabulary to learn (Milton, 2009).

The present review, however, is concerned with what makes words more or less learnable (i.e., with their ease or difficulty rather than their importance). For example, frequent verbs in English which are very often irregular in their forms (e.g., *go: went, buy: bought*) so are in that respect less learnable. Nevertheless, it could be that precisely because these irregular verbs occur frequently, such verbs can still be mastered by learners without too much difficulty. This claim did not become testable until Meara (1992) tested empirically a frequency model which suggested that a word's frequency strongly influences learning it. He produced a graph depicting a frequency profile based on the relationship between frequency and learnability (see Figure 2.2). According to Meara’s (1992) frequency profile, lexical knowledge of a typical L2 learner is highest at the first most frequent 1,000 words and gradually decreases across the less frequent 1,000-word levels.

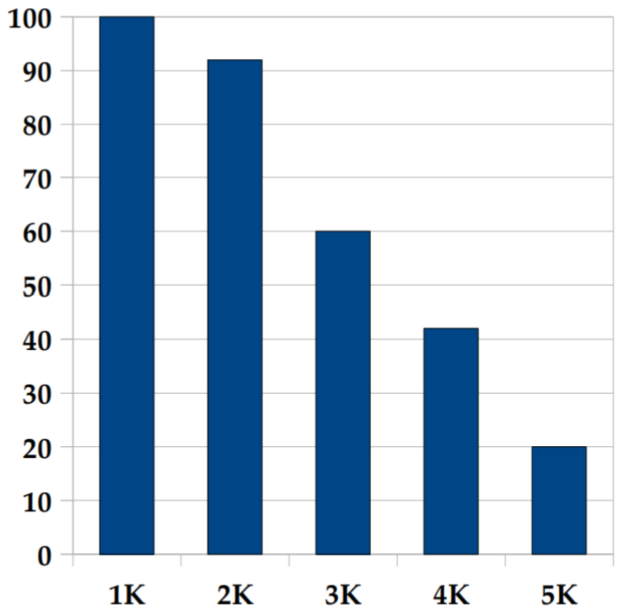


Figure 2.2 vocabulary profile of a typical learner (Meara, 1992/2010, p. 6)

The frequency model has been used in several studies. Milton (2006) tested the frequency model in a study of 227 Greek EFL beginner to upper-intermediate learners via X-Lex (Meara & Milton, 2003). The X-Lex is a computerised test which measures how many words a learner knows from the most frequent 5,000 words in English based on the Nation (1984) and Hindmarsh (1980) frequency lists. Results of the test confirmed Meara's frequency profile, as well as a statistically significant relationship between frequency levels and vocabulary size. Similar findings were observed in relation to learners of languages other than English. For example, the frequency model has been confirmed in learners of French as a foreign language (Richards & Malvern, 2007; Richards, Malvern, & Graham, 2008).

In general, most studies on vocabulary frequency profiles of learners suggest a decrease of known words from the most frequent to the least frequent words, even though the frequencies are drawn from general corpus material and not from text that the learners have actually met in their input.

However, Aizawa’s (2006) study with 363 Japanese EFL learners found that the pattern of the decrease becomes inconsistent and the profile tends to flatten out beyond the 5,000-word frequency level. This probably reflects the fact that the very high frequency words of a language are common in any text, whether it is one that the learner has met or not. Increasingly at the lower frequency levels, however, a word's frequency depends on the topic of the text, which at school in Saudi Arabia, for example, is highly likely to be a Saudi or Muslim topic. Here, then, there is more scope for the frequency in a general corpus to diverge from that in the material that the learner has been exposed to and learned from, so frequency in a general corpus becomes less accurate as a predictor of learning.

In light of this evidence, Milton (2009) cautions that frequency should not be considered alone as a difficulty factor. It is not necessary that frequent words are learnt before infrequent words, because learning vocabulary is often provided in thematic sets and cannot be purely based on frequency. For instance, the days of the week are taught together although their frequencies in corpora often diverge. In addition, arguably, some frequent words are more challenging to learn than infrequent words. For example, learners might find it difficult sometimes to learn how very frequent words combine with, also very frequent, adverbial particles in phrasal verbs.

It should be noted that the effect of frequency on learnability is not confined to corpus-based studies. Several studies have predictably suggested that frequency of occurrence in a foreign language learner’s input (e.g., textbooks, materials prepared by a teacher, teacher’s talk, books, songs) can influence vocabulary learning. For example, Horst, Cobb, and Meara (1998) examined the effect of frequency on incidental vocabulary learning through reading. The study suggested that frequency of occurrence in an authentic reading text better predicted vocabulary learning than overall frequency of occurrence in the language. Particularly, encountering a word eight times and more in a reading text led to sizable vocabulary learning. Similar findings were reported in several studies on incidental vocabulary learning through reading. In general, the number of encounters necessary for vocabulary learning roughly ranges from five to fifteen times (Nation, 1990; Rott, 1999; Webb, 2007).

In sum, frequency of occurrence whether based on leaners' input or an independent language corpus is an important factor that could have an impact on the ease or difficulty of vocabulary learning but it should not be considered in isolation from other vocabulary learnability factors. In the present study, like other factors mentioned in this section, frequency was not used as an independent variable but rather as a variable to be controlled, by holding it constant for all the target items used in the experiment. That is true both for the general frequency of the words in English and the frequency of their occurrence in the materials that the participants are exposed to in the experiment.

2.6.5 Recycling and Repetition

The frequency of words which have been just described, and which impacts on learning, gets created through processes which throughout this thesis have been referred to as *repetition* and *recycling*. In foreign language learning, students come into contact with the frequency generated in these ways through the input they receive and the tasks they perform.

Repetition and *recycling* are not used in exactly the same way, but their difference is quite subtle. *Repetition* is the more common and general word which can include *recycling*, which is a more technical applied linguistic term, in the meaning the current study is concerned with. The following attempt to show the difference through three examples.

First, users of English unconsciously *repeat* words all the time in their speech and writing, this repetition creates the general word frequency recorded in corpora like the BNC and impacts largely implicitly on student learning. This might only rarely be referred to as speakers recycling words.

Second, teachers and textbook writers may ensure that words used in one unit or reading text etc. occur again (i.e., words are *repeated*, in later units or tasks or tests), which creates the word frequency recorded in corpora of textbooks and other material that a learner has actually been exposed to. It may impact on student learning implicitly or explicitly depending on the task where the repetition occurs. This is often referred to as *recycling* done by the teacher, textbook etc.

Third, students may actively create multiple occurrences of words, either autonomously as a learning strategy or prompted by a teacher in a class task, especially when they reread, say aloud, or practice words or other lexical information more than once. The impact on student learning is clearly explicit. This usually refers to in the present study as *repetition* by the learner, not *recycling*. This repetition is of course the way in which students encounter words multiple times in the current study.

Probably the most consistent finding from research into vocabulary learning is that repeated exposures, whether called repetitions or recycling, are essential to maintain vocabulary knowledge (Nation, 2001; Schmitt, 2010). According to Nation (2001), the exact number of the required repeated exposures, which was discussed under frequency above, should not be the main issue. What should matter the most is that recycling and repetition must continue over time and ignoring that might lead to forgetting many partially learned/known words and, eventually, lose all the effort put into learning these words. Once repeated exposure stops, there seems to be a high tendency for vocabulary to be forgotten (Milton, 2008; Schmitt, 2010). Therefore, it is generally suggested to learn new words by recycling and restudying them regularly on a scheduled basis (e.g., Scholfield, 1991; Schmitt, 2010; Al Fotais, 2012).

Furthermore, recycling alone through sheer repeated exposure to L2 input is not particularly effective (i.e., incidental learning), unless some act of remembering words takes place throughout, for example, practicing new words using activities highlighting different aspects of word knowledge (i.e., intentional learning) (Ellis, 1994; Baddeley ,1997; Roediger & Karpicke, 2006).

Both those ideas are reflected in the current study. As mentioned above, the number of repetitions is not the only factor that the current study consider might promote successful learning of vocabulary. The scheduling of the repetitions, and the involvement of explicit learning on each occasion, are regarded as equally important and could have an impact on long-term vocabulary retention. Since, however, the main independent variable in the study is massed versus spaced scheduling of repetitions, the following main section next addresses in detail the literature on scheduling of repetitions and the impact on vocabulary retention of practice with massed and spaced repetitions.

2.7 The spacing effect

General memory research has shown that learners retain information better when learning is spaced, that is, when repetitions are scheduled over a period of time with lengthy intervals between each repetition, as opposed to massed, that is, repetitions in a single lengthy session (e.g., Melton, 1970; Dempster, 1987; Anderson, 1990; Dempster & Farris, 1990; Dempster, 1991; Russo et al., 2002; Seabrook, Brown, & Solity, 2005; Toppino & Bloom, 2002). For example, dividing a fifty-minute learning session to study a word list into five ten-minute learning sessions spread out over a period of time (e.g., days, weeks), is more effective than learning the same word list in one uninterrupted fifty-minute learning session. This phenomenon is called the *spacing effect*, and it is one of the most robust and consistent findings in memory research.

The spacing effect was first observed by Ebbinghaus (1885). As he puts it succinctly, “with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time” (Ebbinghaus, 1885/1964, p. 89). Since then, the spacing effect has been the subject of extensive research over the past decades (for reviews see Pashler, Rohrer, & Cepeda, 2007; Delaney, Verkoeijen, & Spirgel, 2010). For example, the spacing effect has been found to benefit learning of mathematical concepts (e.g., Rohrer & Taylor, 2007), reading skills (e.g., Seabrook, et al., 2005) L2 vocabulary learning (e.g., Goossens et al., 2012), L1 vocabulary learning among children (e.g., Childers & Tomasello, 2002), and learning of English grammatical rules (e.g., Bird, 2011).

Although the spacing effect has been demonstrated in many studies, the reasons behind it are still debated (Delaney et al., 2010). Several theories have been proposed to explain the spacing effect. In general, most theories can be classified into encoding variability theories, deficient-processing theories, or study-phase retrieval theories (Serrano & Muñoz, 2007). The current study does not attempt to investigate whether these theories explain the spacing effect or not. However, a brief review of these theoretical accounts might help shape our understanding of the phenomenon.

2.7.1 Theoretical accounts for the spacing effect

Under encoding variability theories, retention of spaced items is better than massed items because each repetition in a spaced condition is encoded differently, thus, providing more cues for retrieval (e.g., Landauer, 1969; Melton, 1970; Godden & Baddeley, 1975; Glenberg, 1979; Balota, Duchek, & Paullin, 1989). This line of theories suggests that information is usually encoded in memory along with a context (e.g., background noise, events during the study session, etc.) and each context leaves memory traces. Compared to massed learning repetitions, spaced learning repetitions occur further apart in time, thus arguably creating more chances for the context to vary and so make the memory traces and retrieval cues for later recall more diverse.

Deficient-processing theories claim that spaced items are better retained because the length of time between repetitions allows for full processing on subsequent occasions, while massed items do not receive sufficient processing because of the relatively short time between repetitions (Hintzman, 1976; Cuddy and Jacoby, 1982; Challis, 1993). Therefore, in massed learning, the repetitions occur when the first presentation is still relatively fresh in memory, which could mislead learners into paying less attention to the subsequent repetitions because they assume they already know the item better than they actually do (e.g., Bahrick & Hall, 2005).

The study-phase retrieval theories claim that the spacing effect only takes place when the memory trace of the first presentation is not active at the time of the second repetition, thus, the old presentation can be further elaborated upon (e.g., Thios & D'Agostino, 1976). In other words, the spacing effect is more effective because allowing more time between repetitions of an item could actively strengthen its memory traces (Kapler, Weston, & Wiseheart, 2015).

It should be noted that it should not be assumed that only one of these theories explains the spacing effect. On the contrary, the general consensus in the literature seems to lean towards an amalgam of theories to account for the spacing effect (Delaney et al., 2010; Lohnas & Kahana, 2014).

2.8 Research on the spacing effect on vocabulary retention

The impact of the spacing effect on vocabulary learning has been demonstrated in many studies. More than a century ago, Ebbinghaus (1885) conducted a series of systematic verbal vocabulary tasks on himself in a laboratory-controlled setting. He memorized nonsense syllables and repeatedly tested himself over various periods of time. In the first experiment, he memorized a list of nonsense words until he reached perfect recall and then allowed a period of disuse until he was no longer able to recall these items (i.e., 20 minutes, one hour, nine hours, one day, two days, six days, and 31 days). He relearned the same list of items and measured difference in the time saved between learning and relearning. After recording results of these tests, he plotted them on a graph depicting what is now known as the *forgetting curve* (see Figure 2.3). Ebbinghaus concluded that, once exposure to input stops, there is a sharp decline within minutes although relearning items takes less time than learning new items.

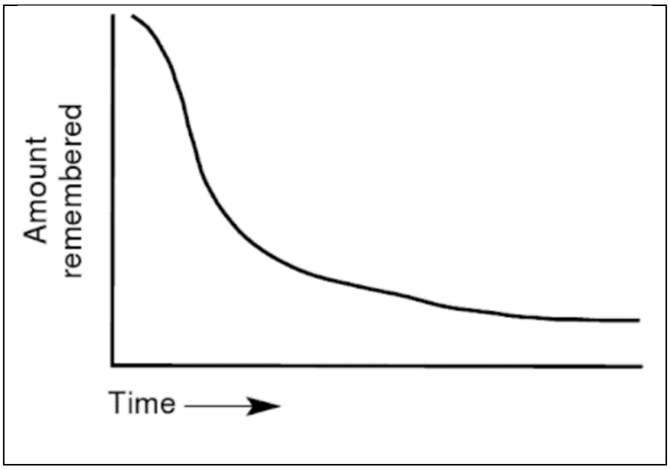


Figure 2.3 typical forgetting curve (Schmitt, 2000, p. 131)

In an early study, Bahrick (1979) investigated the differences between massed and spaced repetition as well as the effect of different lengths of spacing intervals. English native speakers had to learn 50 English-Spanish word pairs in three practice conditions: zero interval repetitions (massed repetition), one-day interval repetitions, and 30-day interval repetitions. The treatment of all groups (i.e., visual paired-associated learning) was followed by six productive recall follow-up tests and the final test took place after 30 days of the fifth follow-up test. Results suggested a preliminary better gain for the massed repetition group; however, the spaced repetition groups showed higher retention later on, suggesting that the spaced repetition technique might help to promote long term retention. After eight years, in a follow-up study, Bahrick and Phelps (1987) were able to track and re-examine 75% of the participants in Bahrick's (1979) study. Results suggested a 2.5-times higher retention among participants from group three (spaced learning distribution) than group one (massed learning distribution). Similarly, Bloom and Shuell (1981) investigated vocabulary spacing effect in high-school students learning French. In two classroom groups, the massed repetition group learned 20 French words (i.e., occupation names) for 30 minutes while the spaced repetition group learned the same words for ten minutes over three days. Results of immediate post-tests suggested similar vocabulary gains, however, the delayed post-test, which was given five days later, suggested higher retention scores for the spaced repetition group than the massed group.

Recent studies have also revealed that spacing can benefit learning vocabulary. In an online study, Kornell (2009) investigated the spacing effect of learning vocabulary with the aid of digital flashcards. Using an online programme, 20 L1 English undergraduate students studied in one learning session 40 English synonyms that consisted of pairing a rare word, which the researcher assumed would be unknown to the participants, and a frequent word that functioned as a definition to the other word (e.g., ‘effulgent: brilliant’). In the massed condition, participants studied 20 synonym pairs in sets of five-word pairs. Each set of word pairs was displayed on the computer screen four times in the same order before introducing the next set. In the spacing condition, all 20-word pairs were presented in a single set (i.e., 20 synonyms) which was repeated four times in the same order. Since all the repetition occurred in one session in both conditions, the spaced condition involved only a very small space of time between repetitions (i.e., the time it took to repeat 20-word pairs) compared with no time delay in the massed condition. The assignment of the order of word pairs per condition was different for each participant. Approximately 24 hours later, an online post-test took place. In a random order, the participants were presented with the first word of each pair and were asked to type down the synonym. Results indicated a retention advantage for items studied in the spaced condition (65% correct responses) over the massed condition (34% correct responses).

In Logan, Castel, Haber, and Viehman (2012), twenty-eight undergraduate students were asked to memorize the form of five words that were presented once, another five words that were repeated immediately (i.e., massed condition) and another five words that were repeated after a lag of three intervening items (i.e., again a very close spaced condition). Participants were informed that they will be studying some words, one at a time, and some of the words might be repeated at some point during the study session. In the single exposure condition, each word appeared on a computer screen for six seconds. After it had disappeared, participants were asked to guess the probability of them later remembering the target word on a rating scale from 0% to 100%. After the learning session, participants were given a distractor task for 30 seconds. After completing the distractor task, a recall task took place in which participants had to say aloud as many (i.e., productive recall of form only) words from the list as possible. Once the recall session was over, the same procedure was followed in the massed and spaced conditions. Results suggested that recall increased with number of repetitions and that there was a spacing effect: participants recalled 67% items in the spaced condition, 51% in the massed condition and 40% in a single presentation condition. In addition, despite better performance in spaced items, participants seem to underestimate the benefits of spaced learning.

More recently, Nakata (2015), echoing Ebbinghaus' forgetting curve, compared the effects of massing, short equal spacing, short expanding spacing, medium equal spacing, medium expanding spacing, long equal spacing and long expanding spacing on L2 vocabulary learning. A total number of 132 undergraduate students studied two English-Japanese word pair lists of ten items each as well as 13 filler items. Participants studied the target words using a computer-based flashcard programme in seven different spacing schedules based on the number of filler items separating each repetition: 0-0-0 (massed), 5-5-5 (short equal spacing), 1-5-9 (short expanding spacing), 10-10-10 (medium equal spacing), 5-10-15 (medium expanding spacing), 30-30-30 (long equal spacing), and 15-30-45 (long expanding spacing). In all learning conditions, each target item was encountered four times. In the first encounter, the English target word and its Japanese equivalent appeared for eight seconds on the computer screen. The rest of the encounters were productive tasks in which participants were required to write down the target English word that correspond to each Japanese translation. All the learning conditions were mixed together. In the massed condition, each item was presented four times in a row, whereas in the spacing conditions, after each encounter, the following encounter is separated by filler items or other target words learned in the other conditions. After the learning session, a filler task took place in which participants answered ten mathematical questions. The task was introduced to minimize the effect of primacy in the post-tests. After the completion of the task, which took around one minute, participants were given an immediate post-test. In addition, a delayed post-test took place one week after the end of the learning session. Results indicated that massing is less effective than spacing when it comes to vocabulary learning. Interestingly, however, results of both immediate and delayed receptive post-test also indicated a statistically significant benefit of expanding over equal spacing. In addition, the study also suggested the effect size and differences in the mean gains were small, suggesting that spacing does not affect vocabulary learning much. These results contradict previous vocabulary study findings in which equal spacing was found superior to expanding spacing and that it had a large positive effect on vocabulary learning (Cull, 2000; Logan & Balota, 2008).

In general, findings from such cognitive psychology research provide some promising results that could be applied to classroom learning (Dempster, 1996; Seabrook et al., 2005). However, it should be noted that most of these studies have investigated the spacing effect based on a single learning session where the spaced condition involves quite short spacing, and/or not in authentic classroom settings. Furthermore, relatively few studies are of L2 vocabulary, and typically only one aspect of word knowledge has been measured in each study. The following section reviews a few studies that have investigated the spacing effect on vocabulary learning in the classroom.

2.9 Classroom-based research into the spacing effect on vocabulary retention

As mentioned in Section 2.5, research on the spacing effect extends to decades and most research into the spacing effect on vocabulary learning shows promising results. However, despite the apparent relevance of the spacing effect to the educational context, most research on the spacing effect was confined to laboratory experiments. There are some possible reasons for the lack of studies conducted in actual classroom settings. In comparison to laboratory studies, there are many confounding variables in classroom studies, which, if not controlled, could affect the success of the spacing treatment (e.g., peer distractions, class session time restraints, loss of participants, etc.). In addition, the delivery of a spacing treatment in a laboratory study is usually computerized. Examining the spacing effect on vocabulary learning using a computer restricts the attention of the participants to the screen, targets them individually and, usually, sets a learning criterion for the treatment to advance from one stage to another. On the other hand, classroom studies are conducted in a class full of peers where the instructor/teacher controls the delivery of the treatment and the attention of students could be diverted for any number of reasons (Kapler et al., 2015).

Only three studies, to the best of my knowledge, attempted to investigate the spacing effect on vocabulary learning in real-world classroom settings. In this section, a brief review of probably the most relevant studies to ours is presented, namely Sobel, Cepeda, and Kapler (2011), Goossens et al. (2012) and Schuetze (2015), which then is followed by a summary of some of the limitations of these studies.

2.9.1 Sobel et al. (2011)

This study examined the difference between spaced repetition and massed repetition in effect on retention of vocabulary knowledge among L1 English fifth-grade students in actual primary classrooms. Thirty-nine students from two intact classes studied just four English words (i.e., two adjectives, one noun and one verb) in the massed condition and another four English words (i.e., two adjectives, one noun and one verb) in the spaced condition. The target words were presented to learners in a booklet, which contained three pages. The items were presented on the first page, the definitions on the second page and the third page was empty to practice writing the target words, their definitions and produce novel sentences. Thus, the form and meaning of the target words were practiced both in receptive and productive modes. The words were different between learning conditions, and there was no controlling of item difficulty.

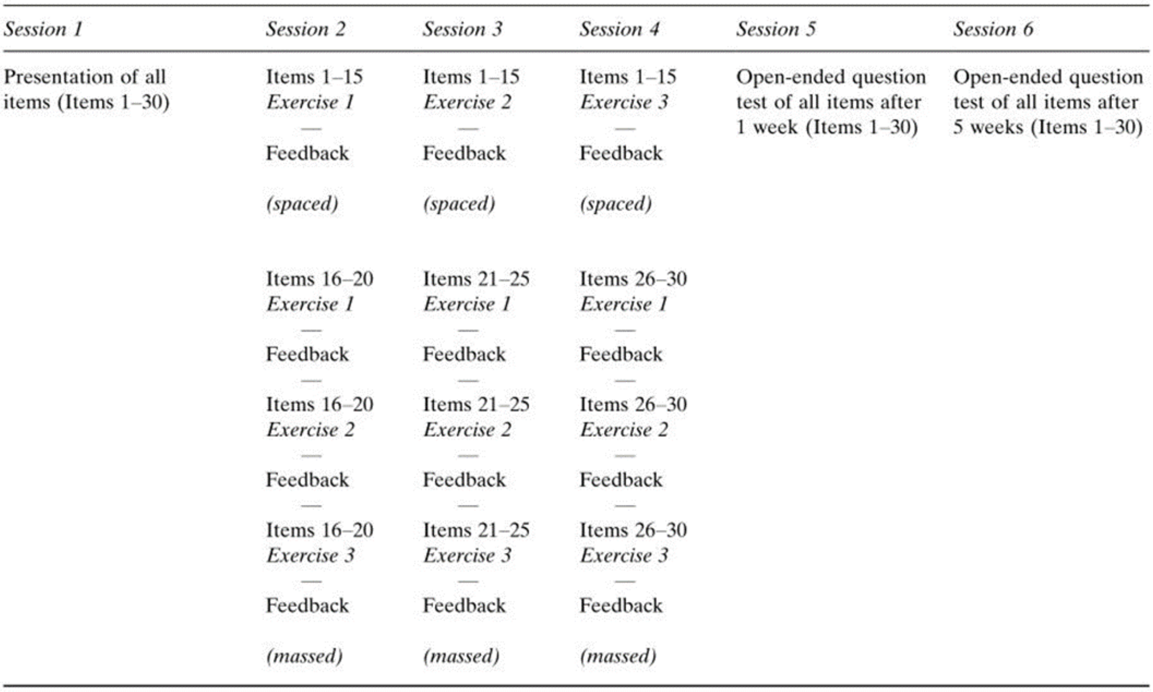
In both learning conditions, the words were studied and practiced using the writing task. However, by the end of the first learning session, the students took a break of less than one minute and restudied the same word list in the massed condition, whereas in the spacing condition, they restudied the list seven days later. A delayed post-test took place after five weeks in which the students had to provide definitions for the studied words (receptive recall). The results indicated a spacing effect: students in the spaced condition successfully recalled definitions of 177% more words than in the massed condition. On average, retention of words learned in the spaced condition was almost three times the number learned in the massed condition.

2.9.2 Goossens et al. (2012)

This study examined the difference between spaced repetition and massed repetition in effect on retention of EFL vocabulary among third-grade Dutch students in actual classrooms. The students learned the meaning of 15 new words through massed repetition and 15 other new words through spaced repetition. The study consisted of four learning sessions, each on one consecutive day followed by a one-week delayed post-test and a five-weeks delayed post-test (see Table 2.4).

In the massed condition, the target 15 words were divided into three sets of five words then each set was taught three times in one of three sessions. All the 15 words in the spaced condition were taught once in each of the three learning sessions. Three types of tasks were performed for each word (i.e., fill-in-the-blank, true/false, and multiple-choice questions). In the massed learning sessions, each word was practiced using all the three tasks, while in the spaced learning sessions, all words were practiced using a different task in each session.

The student retention of the target words was assessed in a receptive recall test one week after the fourth learning session and again after five weeks after the last test. The analysis of the students' responses on the exercises suggested equal good performance in both learning conditions (massed: *M* = 86.81%; spaced: *M* = 87.33%), and no significant difference in performance. However, results of the study found a significant effect of learning condition, F (1, 32) = 10.118, p = .003, η_ρ^2 = .240): retention for words learned in the spaced condition was better than the retention for words in the massed condition. In both post-tests, words learned in the spaced condition were better than words in the massed condition: one-week post-test (massed: *M* = 46.06; spaced: *M* = 55.96), four weeks post-test (massed: *M* = 23.07; spaced: *M* = 27.13). Thus, Goossens et al. (2012) extended the findings of Sobel et al. (2011) that found a spacing effect in vocabulary learning in primary school children.



Session 1	*Session 2*	*Session 3*	*Session 4*	*Session 5*	*Session 6*
Presentation of all items (Items 1–30)	Items 1–15 *Exercise 1* — Feedback *(spaced)*	Items 1–15 *Exercise 2* — Feedback *(spaced)*	Items 1–15 *Exercise 3* — Feedback *(spaced)*	Open-ended question test of all items after 1 week (Items 1–30)	Open-ended question test of all items after 5 weeks (Items 1–30)
	Items 16–20 *Exercise 1* — Feedback — Items 16–20 *Exercise 2* — Feedback — Items 16–20 *Exercise 3* — Feedback *(massed)*	Items 21–25 *Exercise 1* — Feedback — Items 21–25 *Exercise 2* — Feedback — Items 21–25 *Exercise 3* — Feedback *(massed)*	Items 26–30 *Exercise 1* — Feedback — Items 26–30 *Exercise 2* — Feedback — Items 26–30 *Exercise 3* — Feedback *(massed)*		

Table 2.4 procedure of the study (Goossens et al, 2012, p. 968)

2.9.3 Schuetze (2015)

In two studies, Schuetze (2015) examined the difference between the effects of equal and expanding spaced distribution on beginner learning of vocabulary of German as a foreign language at university level. These two studies did not examine massed practice. However, it seems relevant to present these studies in this section because they provide some insights into the implementation of spaced practice of FL vocabulary at university level, so are relevant in that way to the present study.

In the first study, two classes of students formed one group, which followed a uniform spacing schedule, and two other classes of students formed a second group, which followed an expanding spacing schedule. The target vocabulary consisted of 24 content words (i.e., eight nouns, eight verbs and eight adjectives) and 15 function words (i.e., five prepositions, five conjunctions and five adverbs). The wordlist was introduced to the first group using uniform spacing while, on the other hand, the same wordlist was introduced to the second group using expanding spacing. The words were selected from the course's textbook but only words that do not appear in the first five chapters were included to ensure that the student will be exposed to these words for the first time during the first learning session using PowerPoint slides. In addition, words were randomly selected based on word frequency using Jones and Tschirner's dictionary (2006).

In both learning conditions, the words were taught using PowerPoint. On each slide, the L1 (i.e., English) was first presented for two seconds, then followed by presenting the target German equivalent for six seconds. Along with the presentation of each word, a sound was played as well, of a speaker uttering the German word. Each slide lasted for eight seconds and the participants were instructed to write down each German word they saw and heard on a piece of paper (i.e., production recall practice). This continued for the rest of the items, which took about 15 minutes from start to finish. In the uniform spacing condition, the sessions took place on day one, four, eight and eleven. It should be noted that an ideal uniform spaced learning should have an equal interval gap (e.g., FOUR-FOUR-FOUR days interval). However, in this study, the class schedule did not allow a 100% uniform spacing schedule. In the expanding spacing condition, the sessions took place on day one, two, four and eight. In general, each word was presented four times and in each time the order was randomized to minimize order effect.

The second study was conducted a year later. The same procedure as in the first study was followed, with two exceptions: a) two classes of students formed two groups (i.e., 24 students in each group) and b) the relearning sessions were four instead of three. The list was reviewed on day one, four, eight, eleven and fifteen in the uniform spacing condition and on day one, two, four, eight and fifteen in the expanding spacing condition.

Three post-tests were carried out after each study. The first post-test was performed one day after the end of the study, the second post-test four weeks after the end of the study and the last post-test eight weeks after the end of the study. In all tests, the test format was an L1-L2 translation task; that is, the English word was supplied, and the participants had to write down the target German words, so again productive recall. The first study revealed, in all three tests, that the differences between the two spacing conditions were not statistically significant. However, the mean scores for long-term retention in the uniform spacing condition were higher than in the expanding spacing condition. In the second study, the results were similar to the first study. However, increasing the number of repetitions to four times led to a higher recall rate.

2.9.4 Limitations of previous studies

The studies reviewed above seem to indicate that benefits of the spacing effect extend to classroom vocabulary learning. However, it is very difficult to control for all possible factors that might affect the performance of students in classroom studies. The following presents a summary of some of the limitations of previous studies.

First, a vocabulary learning session usually involves using several different tasks. However, only one type of exercise was used in Sobel et al. (2011) and Schuetze (2015), which does not reflect what normally happens in a real classroom learning session. One type of task usually practices only one kind of knowledge of the unknown word. As discussed in Section 2.5, there are many kinds of information to learn about a word and they require different tasks to practice.

Second, and connected with the previous point, the instruments that have been constructed to measure retention at peak attainment in the previous studies were not all fully appropriate to capture the effect of spacing on the retention of more than one aspect of vocabulary knowledge. Each of the three studies mentioned above applied one type of measurement, such as a productive task aimed at assessing the ability to supply the L2 definition of the target words (see further next section). Employing multiple forms of measurement however might yield more fully informative information about vocabulary learning of different kinds of lexical knowledge.

Third, previous knowledge of the target items was not assessed properly. For example, Schuetze (2015) examined previous knowledge of the items only based on a questionnaire that was used to exclude participants who had experience with German, while Sobel et al. (2011) decided that the items were unknown/new to the target participants based on the researchers’ judgment and Goossens et al. (2012) did not examine prior knowledge of the target items at all. It seems imperative in any similar line of research on vocabulary retention to establish a baseline, thus examining previous knowledge of the target items should not be ignored.

Fourth, the number of the items that were selected in the previous studies could arguably be insufficient to produce meaningful and reliable results. Only eight items were selected in Sobel et al (2011), 24 items in Schuetze (2015) and 30 items in Goossens et al (2012). It should be noted that there is no set limit in the literature on how large or small a target lexical sample must be in vocabulary studies, but it is generally assumed that “the more samples you obtain from your participants, the more valid and reliable your results should be” (Schmitt, 2010).

Fifth, the selection of the target item did not consider the possible effect of part of speech on vocabulary learning and/or retention. For example, Sobel et al. (2011) selected four adjectives, two nouns and two verbs, while Goossens et al (2012) selected 25 nouns, four verbs and one adverb. As discussed earlier in Section 2.3.2, word class could affect the ease or difficulty of vocabulary learning and it might be “prudent to control for word class in all vocabulary research” (Schmitt, 2010, p: 160). This could probably be accomplished by targeting one part of speech or by choosing a large and equal number of words with different parts of speech. In addition, word length is also a factor that was not addressed in the previous studies. Schuetze (2015) points out that longer and phonologically similar words were learned the least. However, the small number of long words used in his study did not allow for a statistical analysis.

Lastly, the number of repetitions of the items in the earlier studies is arguably low and does not conform to vocabulary learning literature. In general, the number of repetitions should roughly range from five to fifteen times (Nation, 1990; Rott, 1999; Webb, 2007). However, after initial presentation, the items were repeated once in Sobel et al. (2011) and three times in Goossens et al. (2012) and Schuetze (2015: 1st study). The effect of the number of repetitions was present in Schuetze’s (2015) second study (i.e., increasing repetition from three to four) led to a higher recall.

2.10 Vocabulary testing

As just noted earlier, the testing of vocabulary knowledge is a crucial part of any spaced-massed study, as well as vocabulary pedagogy and research more widely, and is an area of weakness in previous studies due to widespread use of only one measure.

Research on vocabulary testing has been developing hand in hand with research on vocabulary knowledge for the past three decades. Due to the complex nature of vocabulary knowledge and how researchers vary in their view and description of vocabulary knowledge, there is no current comprehensive set of vocabulary measurements that can be used to assess all aspects of vocabulary knowledge of language learners. Instead, there are a few well-established vocabulary tests that can be used on a large-scale to easily assess different areas of lexical knowledge (Milton, 2009).

As stated earlier in Section 2.4, vocabulary growth in terms of breadth positively correlates with linguistic competence; the more vocabulary is known the better the performance in reading, listening, writing, and speaking (Laufer, 1992, 1997; Qian, 2002; Stæhr, 2008; Llach & Gallego, 2009; Mehrpour & Rahimi, 2010). Accordingly, the importance of vocabulary testing arises from the importance of vocabulary itself for language learning. Furthermore, vocabulary breadth or size testing provides researchers and practitioners useful information on "how many words foreign language learners know, how fast their target vocabularies grow, and how these factors are related to other aspects of their linguistic competence" (Eyckmans, 2004, p. 13). Hence in recent years vocabulary size tests have been developed and improved more notably than vocabulary depth tests.

There are certain considerations for choosing or developing vocabulary tests. As discussed earlier, different researchers have different views on vocabulary knowledge. This means that the design of vocabulary tests is subject to the researchers' views on vocabulary knowledge and what aspect of vocabulary knowledge they are interested in examining, and for what purpose. For example, vocabulary tests can be used to assess how many of the target words introduced in a course the language learners have learned by the end of that course (i.e., achievement test), to detect gaps in language learners' vocabulary knowledge and enable language instructors to tackle these gaps more effectively (i.e., diagnostic tests), to assign learners to language learning groups/classes that are suitable for their language level (i.e., placement tests), and can be used as a part of course-independent measures of knowledge (e.g., TOEFL) to give an estimation of learners' overall language skills performance (i.e., proficiency tests) (Schmitt, 1998; Eyckmans, 2004). In the present research, a proficiency measure of breadth/vocabulary size is needed as part of background information about the learners, as well as achievement measures of vocabulary depth for the specific items that the learners have learned in the massed or spaced conditions.

In the rest of this section, I will discuss using multiple measures in testing vocabulary knowledge and review some well-established vocabulary tests to provide background information for my decisions about what measurements to use in examining the vocabulary knowledge of Saudi EFL learners in the present study.

2.10.1 Multiple measures of vocabulary knowledge (depth)

As discussed in Section 2.4, the nature of vocabulary knowledge is rather complex and to examine all aspects of vocabulary knowledge using an all-inclusive and reliable battery of tests at the same time is virtually impossible. Instead, many researchers suggest simply using multiple measures of vocabulary knowledge selectively to give a more comprehensive characterization of vocabulary knowledge and acquisition (Ellis, 2001; Read, 2004; Laufer et al., 2004; Nation, 2007; Milton, 2009; Schmitt, 2010). With respect to depth, it is regarded as good practice for vocabulary studies to use multiple measures of vocabulary which could be accomplished by assessing either "receptive/productive mastery, different types of word knowledge, degree of mastery of an individual word knowledge aspect, contexts of use, etc., or some combination of these" (Schmitt, 2010, p. 22). Yet, as seen in the review earlier, many massed-spaced studies, especially those performed more in the realm of psychology than applied linguistics, have tended to use only one depth measure of vocabulary knowledge of the target words.

There are a number of reasons which justify the need for multiple measures of vocabulary knowledge in vocabulary studies. First, using multiple measures targeting different depths of knowledge makes it more likely that one will include one or more tests "which are more sensitive to degrees of acquisition" (Newton, 1995, p. 171) and so overcome the possibility of a vocabulary test not yielding any usable data (Nation, 2007). It is thus important to consider using multiple test formats that differ from each other in difficulty. On one hand, choosing only one easy format, such as receptive recognition, could overestimate learner' vocabulary knowledge, and perhaps show no signs of acquisition because scores are already high on a pre-test before any experimental treatment. On the other hand, choosing one difficult format could underestimate learner' vocabulary knowledge (Nation & Chung, 2009).

Second, studies that examine learners' vocabulary growth and/or the effect of different vocabulary learning methods on learners' acquisition of vocabulary knowledge usually produce results that could be misinterpreted when using a single test. For example, in four studies, Groot (2000) compared the effectiveness of learning vocabulary using bilingual wordlists against a computer programme for FL vocabulary learning. In study one and two, a receptive translation test was used, and results suggested that wordlists promote better vocabulary learning. However, in study three and four, a productive cloze test was used, and results suggested that vocabulary learning was better when using the vocabulary computer programme. This means that relying on the results of one test would give misleading information about vocabulary knowledge. In the end, Groot drew a conclusion from the results of both measurements and suggested that a combined approach of both learning methods would be beneficial for vocabulary learning. The bilingual wordlists introduce L1 knowledge which assists short-term retention, and the vocabulary learning computer programme employs "intensive processing of the words in the form of the various mental actions" (Groot, 2000, p. 73) which strengthens this knowledge for long term retention. This insight would not have been obtained with only one kind of practice and one kind of test, and this is one reason why the present study pursues multiple practice and test types for the target words.

Third, using multiple measures of vocabulary knowledge would allow for more valid generalizations. In the example of Groot (2000) above, if the study did not use two different measures of vocabulary in the third and fourth study, the results could be mistakenly interpreted as 'generally' meaning that the vocabulary computer programme used in the study is not suitable for vocabulary learning. However, there is also a related issue of whether some kinds of vocabulary knowledge, tested in a certain way, do allow one to infer another kind without testing it. Schmitt (2010) for example comments on the general assumption based on research that productive vocabulary knowledge mastery suggests receptive vocabulary knowledge mastery, because productive knowledge is in some way deeper/more difficult. Hence if one measures production knowledge, one can assume that words that are known on that test are known receptively as well as productively. However, even this is misleading as a productive test alone cannot say which words that *fail* it are in fact known receptively. Furthermore, Schmitt warns that "(t)he real danger is making generalizations in the other direction" (p. 152). It cannot be assumed that receptive mastery ever implies productive as well. Note, however, that even using multiple measurements which elicit receptive and productive vocabulary knowledge separately would not necessarily be sufficient to generally characterise receptive and productive vocabulary knowledge. Other variables are involved in making a test easier or harder such as the recognition versus recall format of test items, discussed next.

As mentioned before, vocabulary learning is an incremental learning process and there are degrees of it even within receptive and productive vocabulary knowledge. Hence each measurement will usually only record vocabulary knowledge at that degree. For example, a multiple-choice test of receptive vocabulary knowledge that measures the form-meaning link at what is often called recognition level, such as where four L1 translations are offered to choose between for an L2 word, can only estimate partial vocabulary knowledge at an early stage of receptive vocabulary learning and the result does not necessarily imply full receptive mastery. By contrast an open response test of receptive vocabulary knowledge that measures the form-meaning link at what is often called recall level, such as where an L1 translation has to be supplied, with no prompt, for an L2 word, estimates fuller vocabulary knowledge at a later stage of receptive vocabulary learning. Thus, while multiple depth measures allow for more generalization, researchers on vocabulary should clearly say what they claim to measure on multiple dimensions (Schmitt, 2010). It is not just a matter of whether receptive versus productive knowledge is tested, but also what aspect(s) of lexical information (e.g., meaning, written form, spoken form, collocation) are targeted, and what level of difficulty is involved in terms of the recognition-recall distinction. In the light of this, the attention of the present study is limited to knowledge of the written form link with one basic meaning for each word and tested both production and reception knowledge of each word at both recognition and recall levels.

In the end, it should be noted that this type of vocabulary testing is complex, however, as it is difficult to define what *vocabulary depth* means and whether vocabulary depth is a distinct dimension of vocabulary knowledge (Read, 2000). For example, it is very difficult to explain the relationship that puts the different lexical aspects such as word associations and functions, knowledge of collocations, and synonyms under the umbrella of vocabulary depth (Milton, 2009). In addition, the number of items in vocabulary depth tests is limited due to extensive testing of different vocabulary aspects which makes the results not representative of the actual vocabulary knowledge of learners across all the words they know (Laufer & Goldstein, 2004).

2.10.2 Vocabulary size tests (breadth)

As already indicated, vocabulary size (breadth) tests give an estimation of the total vocabulary of learners. Usually, this type of vocabulary testing is straightforward as it only measures one low-depth aspect of vocabulary knowledge (e.g., receptive form/meaning recognition) (see Table 2.5). Indeed, vocabulary size tests have been criticized for being limited to measuring superficial knowledge. For example, tests such as the *Yes/No test* (Meara & Milton, 2003) merely ask test-takers to indicate whether they know the meaning of a word or not: however, since their knowledge of the meaning is not tested, this is really only a test of their knowledge that the written word form exists in the language (i.e., a receptive test of knowledge of form alone). Tests such as the

Vocabulary Levels Test (Schmitt, et al., 2001) do ask test-takers to match a word to a definition (Milton, 2009), so test the form-meaning link. However, multiple-choice tests the form-meaning link by offering a range of simple synonyms to choose from, so this is only at receptive recognition level. Consequently, when a vocabulary size test measures one aspect of vocabulary knowledge, it has the advantage of testing a larger sample of items, usually selected to represent frequency bands, which arguably gives a fair representation of learners' total vocabulary size (Read, 2000). However, the depth level chosen is often quite low, which unlike vocabulary depth tests that may measure deeper aspects of learners' vocabulary knowledge for each lexical item (e.g., synonyms, collocates).

Test	Lexical aspect/s
Productive Vocabulary Levels Test (Laufer & Nation, 1999)	Productive meaning→form recall
Vocabulary Levels Test (Schmitt et al., 2001)	Receptive form→meaning recognition
Yes/No test (Meara & Milton, 2003)	Receptive form recall
Computer Adaptive Test of Size and Strength (Laufer & Goldstein, 2004)	Receptive meaning→form recognition Productive form→meaning recognition Receptive meaning→form recall Productive form→meaning recall
Vocabulary Size Test (Nation & Beglar, 2007)	Receptive form→meaning recognition

Table 2.5 Examples of vocabulary tests and the tested lexical aspects

To summarize, vocabulary testing is reliant on how a test designer operationalizes lexical knowledge, what aspect of vocabulary knowledge is under investigation and whether size or depth is assessed. In this study, vocabulary knowledge gained through the practice provided in the massed-spaced experimental conditions is assessed at the basic level of lexical information addressed in most vocabulary tests: the written form - one basic meaning link. Knowledge of collocation, associations, use, lexical grammatical information and spoken form are not addressed. Within that basic information, however, the present study distinguishes four depth levels or 'strengths' defined by the combination of the production-reception and recognition-recall dichotomies. Vocabulary size of participants is also assessed as background characteristic.

The following sections provide reviews of some well-established vocabulary tests which should be helpful to justify my decision when choosing the vocabulary size test used in this study and determining the test(s) that are most suitable to be adapted and/or modified to serve as an achievement test for the main part of the study.

2.10.3 Yes/No tests

Vocabulary size tests designed in Yes/No format typically test random samples of words from frequency bands of the language and ask learners to check *yes* if they know the meaning of a written word or *no* if they do not. This type of test does not ask the learners to provide the meaning of the target words and so relies on the learners’ estimation of their own knowledge. In its traditional form, the testees will soon realise that they are really only being tested on whether they know the wordform exists and do not actually have to know the meaning to respond correctly. Hence its validity as a test of knowledge of meaning at all is doubtful. Furthermore, there is no way to restrict the guessing by learners, which makes the reliability of such yes/no tests questionable, although an adjustment can be made for this by including non-words.

It was Anderson and Freebody (1983) who suggested including non-words in yes/no tests to adjust the overall scores for possible overestimation of vocabulary knowledge. The non-words look like real words and when learners accept one of these words as a known real word, the final vocabulary score is adjusted through a correction formula. In general, most of the formulae that have been proposed to obtain more precise estimates of vocabulary knowledge based on either a simple correction for guessing (*cfg*) procedure (Anderson & Freebody, 1983) or a Signal Detection Theory formula, such as the Δm formula (Meara, 1992) and the Index of Signal Detection (I_{SDT}) (Huibregtse, Admiraal, & Meara, 2002). These scoring methods for Yes/No tests are claimed to provide more precise estimates of vocabulary knowledge. However, each of these scoring methods have some shortcomings. For example, learners' vocabulary knowledge could be underestimated when using the Δm, overestimated when using the *cfg* (Huibregtse et al., 2002), or overestimated only for low proficiency levels learners when using the I_{SDT} (Mochida & Harrigton, 2006). In addition, it is not clear how well the various correction formulae work (Huibregtse et al., 2002; Mochida & Harrigton, 2006) and what percentage of checked non-words must render Yes/No tests invalid (Schmitt, 2010). It should be noted that there is an alternative approach to using a correction formula in scoring Yes/No tests. For example, Schmitt et al., (2011) noted the shortcomings of using correction formulae and decided instead simply to exclude data of participants who checked over 10% of nonwords. That however could damage a study in other ways if a substantial number of the sample of participants gets excluded.

According to Milton (2009), the rate of guessing in yes/no tests varies depending on the culture of the learners. For example, Saudi EFL learners (Al-Hazemi, 1993), Greek EFL learners (Vassiliu, 1994) and Belgian learners (Eyckmans, Velde, Hout, & Boers, 2007) used a significant amount of guessing. On the hand, Japanese EFL learners (Shillaw, 1996) carefully considered their responses in yes/no test and non-words had a little impact on their performance in these tests.

There are several different vocabulary size tests that use a yes/no format; however, three examples of these tests, which have been used to measure the vocabulary knowledge size of Saudi EFL learners, are presented in the next three sections.

2.10.3.1 The orthographic vocabulary size test (X-Lex)

The orthographic vocabulary size test (Meara & Milton, 2003) is a computerized vocabulary size test that measures how many words learners know from the first five 1,000 most frequent words in English. The test is designed in a yes/no format and takes about ten minutes to administer. The test presents learners with 100 real words and 20 non-words. In each frequency level, the test presents learners with 20 randomly selected words and the learners must click on a positive emoticon (e.g., ☺) if they think the word is real or on a negative emoticon (e.g., ☹) if they think it is not. The overall vocabulary size is calculated using the following formula: (correct responses× 50) - (nonwords responses×250).

Similar to other vocabulary size tests designed in a yes/no format, there is a chance that the orthographic vocabulary size test (X-Lex) overestimates the actual vocabulary size of learners. Additionally, there is no limit on the non-word responses that should render the test invalid (Schmitt, 2010). For example, if a learner guessed 100 real words and ten non-words as real words but managed to distinguish the remaining ten non-words, the final score of the learner will be an estimated vocabulary size of 2,500 words. Furthermore, using a positive emoticon instead of *yes* and a negative emoticon instead of *no*, might psychologically influence the decision of the learners. This argument, of course, needs to be confirmed in a study on the impact of using emoticons, representations of facial expressions, instead of the traditional yes/no format on the decisions of examinees. Given these issues, however, this test was not considered in this study.

2.10.3.2 The aural vocabulary size test (A_lex)

The aural vocabulary size test (Milton & Hopkins, 2005) is a vocabulary size test identical in design to the *X-Lex*; that is, it measures vocabulary size within the most frequent 5,000 words in English. However, instead of presenting a word on the screen and asking the learner to decide whether the word is a real word or a non-word, the test plays the sound of the word. Therefore, learners must decide whether the word is a real English word, or a non-word based on its phonological representation. The test does not present the word in writing at any point.

The learners have the chance to replay the sound as much as they like. This is intended to overcome any possibility that the word is misheard. However, in addition to the drawbacks mentioned above regarding the *X-Lex*, the *A-Lex* does not record the replay attempts or take them into account in the final score. This could be useful because repeating the sound of a word many times might indicate that the learner has a little vocabulary knowledge albeit not necessarily of the meaning (Alsaif, 2011). Nevertheless, the *A-Lex* examines the phonological vocabulary knowledge of English, which falls outside the purpose of the present study, so this test was not adopted.

2.10.3.3 Eurocentres Vocabulary Size Test (EVST)

The Eurocentres Vocabulary Size Test (Meara & Buxton, 1987; and Meara & Jones, 1988) was used in Al-Hazmi (1993), which was the first study to assess the vocabulary size of Saudi EFL learners. The EVST is designed in a yes/no format and consists of 150 words representing the knowledge of 10,000 of the most frequent words in English. The test starts assessing the learners' knowledge of the 1,000 most frequent words then moves to words from the next frequency level. In each level, learners are presented with ten real words and another ten non-words. The test asks learners to check *yes* if they know the meaning of the word and *no* if they do not. However, the test stops if scores on one level fall below a predefined threshold. For example, the estimated vocabulary size of a learner is considered between 4,000 and 3,000 words if his/her response on the 4,000-frequency level did not reach that predefined threshold for moving on to the next frequency level.

There are some points that make the EVST not suitable to be used in the current study however. Although it is a practical test to administer, the test stops as soon as a learner shows lack of vocabulary knowledge at a certain frequency level and assumes that it is the same case for subsequent frequency levels. However, some learners tend to show a deficiency in vocabulary knowledge at certain frequency levels and not on other later levels. For example, some studies found notable deficiency of vocabulary knowledge at the 1,000 most frequent words level (Meara & Milton, 2003) and more deficiency of vocabulary knowledge at the 2,000 most frequent words level (Vassiliu, 1994 as cited in Milton, 2009), but better knowledge at later levels. This raises questions regarding the validity of the test as “it presumes that the learner knows even less of the infrequent lexis and does not test it”. (Milton, 2007, p. 52). Yet in many EFL contexts learners have only been exposed to vocabulary in textbooks which may be locally made and not constructed to present words more or less in frequency order based on native speaker English. Additionally, Al-Masrai (2009) in a study that compared between the results of senior and junior university students on the EVST and XK-Lex, a vocabulary size test that assessed vocabulary size knowledge at the first ten 1,000 frequency levels, reported that the EVST might underestimate the vocabulary size of junior EFL learners. The vocabulary size of the junior students on the XK-Lex was between 3,109 and 2,907 words while on the EVST it was 1,680 words.

2.10.4 Multiple-choice tests

This type of test is one of the most frequently used assessment formats in all educational testing. The test format consists of a few options, usually four, and respondents must choose one option that can fill in a gap in a sentence, answer a question, define a word, …etc.

There are several advantages of using this type of testing. The time required to answer and mark this type of test is relatively shorter than that needed for other types of testing, except yes/no tests. Especially with technological advancement and the development of machines that can scan and process data, marking the tests can be automated and done in a short time. Additionally, the results can be instantly available if the test was computerized. Lastly, assessment and marking of this type of testing is objective and marking the test even by hand only usually involves checking the answers against key answers, which could enable any person to mark the test. Two examples of this type of vocabulary test are presented in the next section.

2.10.4.1 Vocabulary Levels Test (VLT)

The Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001) is perhaps the most widely known and used vocabulary size test (Read, 2007; Webb & Sasao, 2013). The Vocabulary Levels Test (VLT) was first designed by Nation (1983, 1990) and has gone through several improvements (Beglar & Hunt, 1999; Schmitt et al., 2001). Although the test was initially designed to help teachers in the development of suitable vocabulary learning materials, as soon as it got published, it became widely used internationally in English speaking countries to test the vocabulary size of international students and migrants at a range of vocabulary frequency levels (Xing & Fulcher, 2007).

The VLT is designed to produce the vocabulary profile of learners at five frequency levels (i.e., 2,000-word level, 3,000-word level, 5,000-word level, Academic word Level and 10,000-word level). The word frequency levels are based on Thorndike and Lorge’s word list (1944), the General Service List (West, 1953) and Francis and Kucera’s word list (1967).

The test is a type of multiple-choice test and requires respondents to choose the meaning of a written English word from a list of possible answers also in English. In its latest version (Schmitt, et al., 2001) the VLT contains 30 items at each frequency level and comes in three versions (B, C and D). Unlike traditional multiple-choice tests, the VLT minimizes guessing by introducing the items in clusters of six words and three definitions. In addition, the words in each cluster are always put in an alphabetical order, as shown in the example below:

1. business
2. clock _____ part of a house.
3. horse _____ animal with four legs.
4. pencil _____ something used for writing.
5. shoe
6. wall

The VLT has been shown to be a valid and reliable test (Beglar & Hunt, 1999; Schmitt et al., 2001). However, the VLT might not be an adequate vocabulary size test to be used in this study due to the following limitations. First, the frequency lists that were used for creating the VLT are very old and there might be "variation between the occurrence of words 50-70 years ago and today" (Webb & Sasao, 2013, p. 265). For example, the 2,000-word frequency level was created using the General Service List (West, 1953), whereas the 3,000, 5,000, and 10,000-word frequency levels were created using Thorndike and Lorge (1944) and Kuĉera and Francis (1967). Second, each question in the VLT consists of three test items and six possible choices. Thus, it is possible that "the learner's knowledge of some of the items is likely to have an impact on the ability to work out the answers to other items where these are not known" (Milton, 2009, p. 75). Lastly, the VLT "is not really designed to provide an estimate of a person's overall vocabulary size [...] the test is better used to supply a profile of learner's vocabulary, which is particularly useful for placement and diagnostic purposes" (Schmitt, 2010, p. 198). Furthermore, if the test is used to give the total vocabulary size, it produces an estimation of vocabulary size based only on four-word frequency levels (i.e., 2nd 1,000 frequency level, 3rd 1,000 frequency level, 5th 1,000 frequency level, Academic word Level and 10th 1,000 frequency level). This means that the test uses the knowledge of the 2nd 1,000-word frequency level to estimate vocabulary knowledge of the 1st 1,000-word frequency level and the 5th 1,000-word frequency level to estimate vocabulary knowledge of the 4th 1,000-word frequency level. In Al Fotais (2012), the vocabulary size of Saudi EFL learners was 1,447 words out of the 1,000 to 3,000-word frequency levels and around 48% of the known words came from the 1,000-word frequency level. Accordingly, the participants in this study (i.e., Saudi first year English major students) are expected to have a similar vocabulary size. Therefore, the VLT might not be suitable to measure the vocabulary size of the participants in this study as it does not effectively examine vocabulary knowledge at the 1,000-word frequency level.

2.10.4.2 Vocabulary Size Test (VST)

The Vocabulary Size Test (Nation & Beglar, 2007) is a written four-option multiple-choice receptive vocabulary size test. It consists of 140 items measuring vocabulary knowledge at fourteen 1,000 spoken word family bands in the British National Corpus (BNC) (Nation, 2006), from the first 1,000 to the 14th 1,000-word frequency level. At each frequency level, ten items were randomly selected, so each item represents 100 word-families within the same word frequency level. The following is an example from the first 1,000-word family frequency level:

period**:** It was a difficult **period**.

a. question

b. time

c. thing to do

d. book

The VST is claimed to "provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1,000 to the 14th 1,000-word families of English" (Nation & Beglar, 2007, p. 9). However, there are a few possible shortcomings that should be noted. In a recent case study, Gyllstad, Vilkaite and Schmitt (2015) examined the effect of guessing and sampling rate on data from the VST. The study compared test-takers' performance on the three sections of the VST (i.e., 3K, 6K, and 9K sections) with their performance on follow-up interviews where participants were asked to look at a list of words from each section of the VST without the aid of multiple-choices and describe the meaning of those words (e.g., L1 translation equivalent, L1 or L2 definition, L2 synonym). The results suggested that there was a significant difference between the participants' scores in the VST and the interview on the 3K and 9K sections: there was a tendency for the VST to overestimate the participants vocabulary size at these frequency bands.

The researchers provided two possible explanations for the difference between test-takers' scores on the VST and the follow-up interviews, where scores were lower. First, the VST requires test-takers to demonstrate knowledge at a less demanding level (i.e., receptive meaning recognition), whereas the oral interviews elicited knowledge at a more demanding level (i.e., productive meaning recall). This explanation is congruent with Laufer and Goldstein's (2004) finding that meaning recall tasks are more difficult than meaning recognition tasks and it does not represent a deficiency in the VST, just that it was not testing the same level of depth of knowledge as the interview did.

The second explanation for the discrepancy in scores between the VST and the interviews could be a result of overestimation due to guessing, however. The researchers therefore highlighted the issue of a clear overestimation tendency in multiple-choice tests and called for a careful consideration when choosing multiple-choice tests, such as the VST, as a vocabulary measurement instrument for pedagogical and research purposes.

In addition, in respect to the sampling rate, the VST consists of ten target items for each word frequency level with a ratio of one target item to 100-word families at the corresponding word frequency level. Therefore, each target item's "characteristics (e.g., cognate, or false friend or not) and each test item's efficacy (strong or weak item) has a disproportionate effect on the overall vocabulary size estimate" (Gyllstad, et al., 2015, p. 280). The results suggested that the VST sampling rate of one target item to 100-word families is roughly sufficient. However, the researchers suggested that due to the possibility of overestimation, the VST might not be suitable in situations when accurate vocabulary size estimate is required (e.g., estimating graded reader levels). It should be noted that Gyllstad, et al (2015) did not arrive at a conclusive answer to the difference between scores on the VST and oral interviews. In my view, any or both of the above-mentioned explanations could be possible and the effect of guessing on vocabulary size multiple-choice tests should be further examined.

Despite the possible limitations of the VST, it is the most suitable vocabulary size test for the participants in the current study for several reasons. Most importantly, the VST provides an estimation of vocabulary size knowledge at each word frequency level from the 1st 1,000-word frequency level to the 14th 1,000-word frequency level. The range of word frequency levels that the VST covers is important in this study for two main reasons. First, the participants in this study are expected to have a low vocabulary level that mostly consists of words form the highest word frequency levels (i.e., 1,000-word frequency level and 2,000-word frequency level). Secondly, the target items in the study were selected from a lower word frequency level to minimize the chances of these items being known by the participants (i.e., 5,000-word frequency level). Therefore, examining the participants' vocabulary size at the same word frequency level as that of the target items, as well as at lower word frequency levels, would further confirm the expected vocabulary size of the participants and support the decision to choose the items from a low word frequency level. Second, in terms of classroom time constraints, the test format makes it practical in the target context as it would take a relatively short time to administer.

2.10.5 Translation tests

This type of testing usually assesses vocabulary knowledge by asking learners to supply the translation of a given lexical item, although multiple choice versions are also possible, where a range of L1 or L2 forms is offered. Compared to other types of vocabulary testing (e.g., multiple-choice test), the open response format of translation tests that require learners to supply the translation of a word does not provide learners with any chance of guessing. Furthermore, open response translation tests provide accurate comparisons between receptive and productive vocabulary knowledge because the test format can be identical (i.e., recall: receptive knowledge L2→L1, productive knowledge L1→L2) (Webb, 2008).

While this type of testing seems more appealing in the study of vocabulary, its major drawback is that it cannot be easily used in a situation where the L1 of all testee s is not uniform. Therefore, it might be possible to use this type of testing in an EFL situation where learners share the same L1 but would be infeasible to use in an ESL situation where learners usually have different L1s. An example of a vocabulary translation size test is provided in the following section.

In this study of course the L1 of the participants is uniform, as they are all L1 Arabic speakers. I did therefore make some use of translation tests, both open response and multiple choice, not for estimating vocabulary size but, in the testing of depth of knowledge of the targeted vocabulary of the experiment.

2.10.5.1 Computer Adaptive Test of Size and Strength (CATSS)

CATSS (Laufer and Goldstein, 2004; Laufer et al., 2004) is a computer-based online vocabulary size test (of basic form-meaning pairings) that aims to assess vocabulary knowledge at four levels of depth/strength. The levels of strength are made up from crossing two oppositions which were described earlier: recognition vs recall (operationalized as ability to choose from multiple-choice alternatives versus to supply the answer) and mode (receptive and productive knowledge). These levels are, in increasing strength: *receptive recognition*, *productive recognition*, *receptive recall*, *productive recall*.

CATSS uses the frequency levels that were established by Nation (1983), i.e., the 2,000 most frequent words, the 3rd 1,000 most frequent words, the 5th 1,000 most frequent words, the 10th 1,000 frequent words and the AWL. However, instead of using 18 words at each level, CATSS uses 30 words. Because each word is tested in four degrees of strengths, the total number of the test's items is 600. However, the computer adaptiveness that CATSS utilizes does not mean that each learner will be presented with all 600 items. At each frequency level, the test begins testing the learners' vocabulary knowledge at the productive recall level, which is assumed to be the hardest level of vocabulary knowledge. If a learner's response to a word is accurate, the test moves on to the next word and does not test the same word again at the other remaining strengths (i.e., *receptive recall, productive recognition, and receptive recognition*). If, however, a learner does not know the meaning of the word or responds inaccurately, the test keeps the word in its memory for a later presentation at other strength levels of word knowledge. It should be noted that the word that received an inaccurate response is not re-presented immediately but rather after all words from the same word frequency have been tested at the same strength level of knowledge. Once an accurate response is recorded the test does not present the word on the next strength level of word knowledge. After the test at all four strength levels has been completed in this way at one frequency band, the test moves on to the next word frequency band and its four strength levels of word knowledge. The test therefore *assumes* that the four levels of word knowledge form a strict implicational scale so that if a word is known at a stronger/deeper level then it is inevitably also known at a weaker/shallower level and need not be tested again at that level.

At the end of the test, a result sheet appears showing the overall vocabulary size, how many words a learner knows, as well as the vocabulary size and strength score, how many words a learner knows and the strength of this knowledge. Unlike most vocabulary size tests which usually examine word knowledge at one depth/strength level (e.g., *receptive recognition*), CATSS takes into account the differences of the strength of word knowledge. In addition, the test appears to be reliable on all levels of strength.

Laufer et al., (2004) investigated the validity of the monolingual version of CATSS based on three expectations: high frequency items are easier to learn than low frequency items, there is a hierarchy of vocabulary knowledge strengths, and that there is a relationship between size and strengths of vocabulary knowledge. The test was administered in pencil-and-paper format to a sample of intermediate to advanced L2 university learners in Australia. In order to minimize any contextual effect, the participants were given two strength tests in one setting and the other two in another sitting a week later. The participants were given all 30 items at each strength level and also progressed through all word frequency levels from the 2,000 to 10,000.

Concerning the first assumption, the test indicated that second language learners were indeed more likely to have learned high frequency words before low frequency words, so this supported validity of the test. Furthermore, results validated the second assumption that there is a strength hierarchy of vocabulary knowledge in the form of productive recall being more difficult than receptive recall, followed by productive and receptive recognition, respectively. Implicational scaling, however, only supported three rather than four strength levels due to productive recognition and receptive recognition being similar in difficulty. Productive recognition and receptive recognition were not distinguishable in terms of strength due to similarity between the test formats: choosing from four options the correct word form that matches a target definition was not more difficult than choosing the correct definition for a target word (Laufer et al., 2004).

As for the third assumption, results suggested indeed that size and strength of vocabulary knowledge are related constructs and the estimate of vocabulary size may be lower on more demanding strength levels (i.e., productive, and receptive recall) and higher on less demanding strength levels (i.e., productive, and respective recognition).

In another study, Laufer and Goldstein (2004) investigated the validity of the hierarchical assumptions concerning the four degrees of strength in the bilingual version of CATSS and the relationship between the different degrees of strength and academic success. A total of 435 high school and university L2 learners were given CATSS in a pencil-and-paper format. The L1s of the participants varied as 278 participants were native speakers of Hebrew, 140 participants were native speakers of Arabic, and 17 participants were native speakers of Russian. The L1 speakers of Arabic took an English-Arabic test, while the L1 speakers of Russian were L2 Hebrew speakers so they took the Hebrew-English test. The participants were divided into groups and each group was tested on one frequency level: 9th and 10th graders were tested on the 2,000 words frequency level; 11th and 12th graders were tested either on the 3,000-word frequency level, the 5,000-word frequency level, or the Academic Word list level; the university students were tested on the 5,000 words frequency level. Due to the low proficiency level of participants, the 10,000-word frequency level was excluded from the study.

Each subtest consisted of 30 items and the order of the items in each subtest was randomized to minimize practice effect. Two subtests were administered in one class session and another two subtests were administered in the following class session. The participants were first given the productive recall subtest then, upon completion and collection of the test, half the participants were given the receptive recall test, and the other half the productive recognition test. Once this subtest was completed, the test sheets were collected and the participants who received receptive recall earlier were given the active recognition test, while those who received the productive recognition test earlier were given the receptive recall test. Once the participants completed this subtest, the test sheets were collected, and the last subtest was given to the participants (i.e., receptive recognition).

The results of Laufer and Goldstein’s (2004) study supported the hypothesis that the word form-meaning link is known at four distinct levels of knowledge in a hierarchy of four levels of strength, with the lowest scores achieved at the hardest strength level (i.e., productive recall) and highest scores achieved at the easiest level (i.e., receptive recognition). Furthermore, results strongly supported an implicational scale where the sequence of the four strength levels was confirmed at all word frequency levels. In the monolingual version of CATSS (Laufer et al., 2004), productive recognition and receptive recognition were indistinguishable in terms of strength, while in the bilingual version, receptive recognition was easier than productive recognition, as predicted. According to Laufer and Goldstein (2004), the difference between recognition results in the monolingual and bilingual versions could be attributed to the receptive recognition subtest in the bilingual version (i.e., choosing the correct L1 translation of the target word) being easier than the receptive recognition subtest in the monolingual version (i.e., choosing the correct L2 definition of the target L2 word). Consequently, in the bilingual version, receptive recognition was significantly easier than productive recognition. Despite the differences between the monolingual and the bilingual versions of CATSS, results validated that "the ability to recognize words, whether passively or actively, will generally precede the ability to recall them, and that recall of meaning will precede the recall of form" (Laufer & Goldstein, 2004, p. 423)

Regarding the relationship between the different degrees of strength and academic success, each strength level correlated significantly with the participants' English class scores, especially at the receptive recall level (.65). Finally, result of the regression analysis indicted that knowledge of form-meaning link could explain 42.6% of the variance in the participants' class grades.

Although CATSS addresses some limitations of other vocabulary size tests, it might not be suitable to be used as a vocabulary size test in the current study for the following reasons. First, the monolingual version of CATSS is probably more suitable for advanced EFL learners. The participants in the current study are first-year English major university students, which means that they will most likely have to answer each word in more than one modality. Therefore, it is expected that it would take a significant amount of time for the test to be administered. Second, the bilingual version of CATSS comes in a Hebrew-English format and developing an Arabic-English version would take time and effort which extends beyond the scope of this study.

However, such an instrument does seem suitable in a number of ways for constructing an achievement test as required for testing the targeted vocabulary in the experiment itself. The bilingual version of CATSS is designed to assess both the size and four levels of strength/depth of vocabulary knowledge. According to my knowledge, as shown in Sections 2.8 and 2.9, most previous vocabulary research that tackled the phenomenon of the spacing effect relied only on one form of measurement at one depth/strength level. Therefore, using an achievement test based on a tried and tested format that examines the strength of vocabulary knowledge at four levels is a valuable contribution as it could help shed light on the differences between the two learning conditions in better detail.

Second, the levels of strength covered by CATSS adequately reflect the depth levels of the types of exercise commonly used by Saudi EFL learners. The four levels tested by the CATSS all remain within the realm of basic word information (i.e., the written form and one basic meaning) that is predominantly the scope of exercises that Saudi EFL learners employ in learning new vocabulary (Alhatmi, 2012, Alharthi, 2012). Therefore, it makes sense for the current study to remain within the scope of those four CATSS levels of word knowledge depth. Therefore, I will adopt exercises for the practice/repetition phase of the study that target those four CATSS levels and use similar tests to those of CATSS for the measurement of word knowledge.

2.11 Conclusion

Vocabulary learning is essential in learning a foreign language. In general, vocabulary knowledge could be described in terms of the ability to recognize a word in reading or listening (receptive knowledge) or the ability to use a word in writing or speaking (productive knowledge). Furthermore, vocabulary knowledge could be described in terms of the amount of words that learners know (size of knowledge) or how well they know these words (depth of knowledge). The current study primarily addresses depth. It assesses changes in the receptive and productive vocabulary knowledge of first year EFL Saudi university students at two further strength levels, and how that all is affected by two learning conditions. It also assesses their receptive vocabulary size.

It is well established that vocabulary learning is incremental in nature and it requires multiple exposures to maintain vocabulary knowledge. A robust finding in memory research suggests that scheduling vocabulary repetitions over a period of time with some intervals between each repetition helps maintain vocabulary knowledge better than repetitions in one lengthy session. Although research on the spacing effect goes back several decades and it has obvious relevance to the educational context, most research on the spacing effect has been confined to laboratory experiments. Studies that have examined the spacing effect in vocabulary learning were not conducted under classroom conditions (e.g., Miles & Kwon, 2008; Nakata, 2008; Kornell, 2009; Kapler, et al., 2015). Only a few recent studies, to my knowledge, have investigated the spacing effect in vocabulary learning in real world classrooms, namely Sobel, et al. (2011), Goossens et al. (2012) and Schuetze (2015).

In the field of vocabulary assessment, different tests have been proposed. Reviewing some important and widely used vocabulary size tests in the literature, I have decided in the current study to use the VST (Nation & Beglar, 2007) as the measuring tool for receptive vocabulary size knowledge. The test is believed to "provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the first 1,000 to the 14th 1,000-word families of English" (Nation & Beglar, 2007, p. 9).

Moreover, the instruments that were constructed to measure retention of specially taught words at peak attainment in the previous massed-spaced studies were not deemed appropriate in various ways to capture the effect of spacing on the retention of vocabulary knowledge. For example, Sobel et al (2011) and Goossens et al (2012) applied only one type of measurement, which was a receptive task aimed at assessing the ability of participants to supply the L2 definition of the target words. It is reasonable to propose that employing multiple forms of measurement of knowledge at different strength levels might yield more informative results. Therefore, I have decided to construct an achievement test based on the format of the bilingual version of CATSS to assess vocabulary retention after repetition with massed and spaced distribution in four strength modalities, namely productive recall, receptive recall, productive recognition, and receptive recognition.

Finally, it should be noted that in the field of EFL in Saudi Arabia, no previous research has tackled the issue of the spacing effect in the classroom. Therefore, the current study aims to shed much needed light on a very important aspect of vocabulary learning.

2.12 Research questions and hypotheses concerning effects of spacing on vocabulary learning

Based on the review of spacing effect studies above, there seems to be a necessity to posit the following RQs and Hs:

RQ 1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?

H1: Massed and spaced practice have equivalent effects on vocabulary learning in the form of initial vocabulary knowledge gains (e.g., Bloom & Shuell, 1981). It should be noted that no previous attempts have been made to investigate the difference between massed and spaced practice in effect on vocabulary learning under real classroom conditions in immediate post-tests (i.e., Sobel et al., 2011; Goossens et al., 2012).

RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test?

H2: It is expected that spaced practice leads to higher gains in vocabulary knowledge on delayed post-tests than massed practice. Research by Sobel et al. (2011) and Goossens et al. (2012) revealed better long-term recall of lexical items when learning opportunities were spaced over several learning sessions as opposed to massed in one single learning session.

The following RQs were formulated without corresponding Hs since no previous attempts have been made to investigate the spacing effect on certain aspects of vocabulary knowledge (i.e., word class, word knowledge strength), or on EFL learners' perceptions of the benefits of spacing vocabulary learning and the relationship between EFL learners' favouring of spacing/massing and their vocabulary achievement scores on the immediate and delayed post-tests. Consequently, the current study attempts to answer the following questions:

RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength level between the immediate post-test and the delayed post-test?

RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed?

RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed vocabulary learning?

RQ 5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test?

Chapter Three - Methodology

3.1 Introduction

As described in Section 2.9 in Chapter 2, earlier studies investigating the effect of massed and spaced practice on classroom vocabulary learning have suggested that spaced practice leads to a higher degree of long-term learning than massed practice. Sobel et al. (2011), for instance, examined the difference between the effect of spaced practice and massed practice on retention of English vocabulary knowledge among native speaker US fifth-grade students in actual primary classrooms. The level of knowledge tested was receptive recall (L2-L1 translation test). Scores which emerged from the study suggested that retention of words learned in the spaced condition was almost three times that of words learned in the massed condition.

Goossens et al. (2012) also examined the spacing effect, albeit in English vocabulary learning by Dutch children, at primary school level, using more words than Sobel et al., and targeting the productive recall level of knowledge. The study concluded that massed and spaced practice led to equally good performance on a one-week delayed post-test while, on the other hand, spaced practice led to better performance than massed practice on a five-weeks delayed post-test.

However, both Sobel et al. (2011) and Goossens et al. (2012) had design weaknesses. The former began with one part of the spaced treatment and all the massed treatment during the same learning session, followed in a later session by the second part of the spaced treatment. This design might arguably have provided words learned in the spaced condition with some memory benefits due to being practiced last, so having the benefit of recency effect when the test was taken. Goossens at al. (2012) used a more sophisticated design where in each of three sessions both spaced and massed learning occurred, but the former was always done first. That meant that some of the words learned in the massed condition had the final recency advantage while the spaced words had the general advantage of higher student concentration levels. Furthermore, neither of these studies measured participant knowledge of the target vocabulary before the study, to establish a baseline. Hence, they use the scores only from tests after the practice treatments as an indication of learning, which assumes a baseline level of zero knowledge, which may not have been entirely the case[5].

In addition to the design weaknesses of previous studies in classrooms, there are also gaps in the aspects of the spacing effect which have been studied. As remarked in Chapter 1, to my knowledge, no earlier research examined the spacing effect in EFL classroom vocabulary learning at university level. Although the research suggests that long-term vocabulary learning is better when learning is spaced over multiple sessions than when learning is massed in one session, it is not exactly clear how much vocabulary is learned and when, as well as if the advantage of spacing over massing in classroom vocabulary learning is equally beneficial to university level learners. Furthermore, there has been no systematic account of how these treatments may affect learning of vocabulary differentially at different depth/strength levels of lexical knowledge. The present study, therefore, will systematically investigate the spacing effect in EFL vocabulary learning at university level to measure how much target vocabulary the students have learned at four different knowledge levels, immediately and four-weeks after the practice phase.

Details of the study's objectives, subjects, design, materials, instrumentation, and procedure are explained in the following sections.

3.2 Objectives

The current study aims to achieve the following objectives:

1. to investigate the relative effectiveness of massed practice and spaced practice in short-term vocabulary learning. It will check vocabulary achievement scores at the receptive

[5] Limitations of earlier studies were reviewed in Section 2.6.4.

recognition, productive recognition, receptive recall, and productive recall levels of knowledge.

2. to investigate the relative effectiveness of massed practice and spaced practice in long-term vocabulary learning. It will check vocabulary achievement scores at the receptive recognition, productive recognition, receptive recall, and productive recall levels of knowledge.
3. to investigate the attrition of vocabulary knowledge through comparison of the immediate vocabulary achievement test scores of the participants with the delayed vocabulary achievement test scores. It will compare massed and spaced practice scores at the receptive recognition, productive recognition, receptive recall, and productive recall levels of knowledge.
4. to investigate the effect of word class on learning, through scores in the immediate and the delayed vocabulary achievement post-tests. It will compare massed and spaced practice scores for learning nouns and verbs at the receptive recognition, productive recognition, receptive recall, and productive recall levels of knowledge.
5. to investigate the participants' perceptions of learning vocabulary using massed practice and spaced practice.

3.3 Research paradigm

Consistent with the prevailing research paradigm used in this field in the studies reviewed in Chapter 2, it is possible to describe the current research as predominantly positivist in nature (e.g., McNamara, 2001). It approaches the issue at hand very much top down, with clear and highly specific research questions posed in advance, which will be answered quantitatively. It could also be seen as, in effect, testing the hypothesis, derived from most of the literature that was reviewed, that massed practice will be less effective than spaced practice. Further hypotheses tested include the common finding that nouns are easier to learn than verbs and the expectation that achievement scores will fall at successively higher levels of vocabulary knowledge.

Consistent with the positivist paradigm, attempts are also made: to sample defined populations of participants and words to be learned; to control key variables such as age, gender, educational background of participants, and their prior knowledge of the words tested; and to follow a recognised and effective research design.

Nevertheless, there are some features of the study that are typical of other paradigms. First, the questionnaire did contain an open response question which allowed for qualitative data to be gathered on any aspect of the issue at hand which participants wished to comment upon. This feature, therefore, allows for issues to arise which were not determined in advance, but rather come from the participants' own perceptions and experiences. That is typical of the constructivist approach to vocabulary research (e.g., Giridharan, 2010) which moves bottom-up from data to theory, rather than the reverse.

Finally, it was also an aim of the study to discover information that would be relevant to language teaching in real EFL classrooms, particularly in the context where the study took place. Hence the study took place in that target context rather than in an artificial environment such as a laboratory (e.g., Seabrook et al., 2005; Kornell, 2009; Logan et al., 2012). Furthermore, the current study used real words that the students might as well meet later in their normal classes, and I was at the same time a researcher and the teacher of the students. Such characteristics are not typical of the positivist paradigm which emphasises detachment of the researcher from the research situation to ensure objectivity, and the use of artificial stimuli and situations where there can be maximum control of the conditions in which the data is gathered (McKinley, 2016). These characteristics are, however, consistent with the action research paradigm whose aim is particularly to intervene in a real situation of which the researcher is a part, while maintaining everything as naturalistic as possible, in order to test the benefit of some new 'action' to improve the performance of the participants (Wallace, 1998; Mackey & Gass, 2013).

3.4 Main study participants

The participants of the present study were 62 Saudi first year English major students at Taif University in Saudi Arabia. It should be noted, that all the participants were male because the higher educational system in Saudi Arabia is gender-segregated, making it impossible for a male researcher to access female students in the ways that the present study necessitated. Furthermore, the age of the participants ranged from 19 to 20, with 59 participants being 19 years old. Most participants had begun formal English learning in schools around the age of 13, which means that they had at least six years of previous English study. The estimated vocabulary size of the participants, as measured by the VST, ranged from 800 to 2,100-word families (*M* = 1408.2, *SD* = 296.4).

The participants were recruited from two intact classes; class one consisted of 33 participants while class two consisted of 29 participants. The participants in both classes were enrolled in a ‘Reading Skills’ module which was a first semester first year course. This course aimed at developing students’ reading comprehension, summarizing, paraphrasing and vocabulary building.

The students knew that they were participating in a study, read participation information sheets and signed consent forms. The participants were informed that, while course attendance and participation were mandatory, they had the right to opt out from including their questionnaire and test data in the study. Additionally, the participants were informed that failure to participate would not reflect negatively on their grades for the actual taught courses, however, they would receive ten credit points in the taught course for full participation in the study. All the students from both classes agreed to take part in the study.

3.5 Design

The study presented here examined the spacing effect in a real classroom setting. Therefore, it is a quasi-experiment as the participants were in intact classes. The participants were not selected randomly from the population of first year students, nor divided randomly between the two classes. In general, the random assignment of participants to experimental conditions is viewed “as one of the hallmarks of experimental research” (Mackey & Gass, 2013, p. 146), because it helps control confounding factors such as individual differences. However, a quasi-experimental design such as in this study reduces threats to ecological validity, which I regard as crucial for the current study. A natural classroom environment best represents actual educational settings and, thus, provide a license later for the researcher to use the study’s findings to inform teachers about how best to improve the vocabulary learning of their students in the classroom.

Furthermore, the lack of true randomness of participant selection was largely mitigated by the fact that the design of the study was not between groups but within subjects/repeated measures. Since all the students in both classes experienced both the practice treatments (massed and spaced learning), their individual differences were present in both conditions and could not be confounded with the treatments. Although different words had to be used as learning targets in the two treatment conditions, a counterbalancing method was employed. The words learned in the massed condition in one class were learned in the spaced condition in the other, while the reverse occurred for a second, equivalent, set of words. Hence, to a great extent, any shortcomings due to the non-random assignment of participants to classes were overcome.

A further feature of the study with potential dangers was the use of real words for the students to learn. Again, using real words was preferred for realism, so that the study resembled as closely as possible real classroom learning conditions. However, it carried the danger that some of the words would be already partly known by participants, which could only be eliminated by the use of unreal words, or words from an obscure language (e.g., Yupik languages). Therefore, the students in the current study participated in pre-tests of the target words one week prior to the study (see Section 3.4). This method ensured that none of the participants had previous knowledge of the target vocabulary which was finally selected.

In order to address some of the other possible issues, participants from both classes completed a background questionnaire one week prior to the study (see Section 3.4). The questionnaire was administered to ensure that none of the participants suffered from memory or medical problems, part or full-time job commitments while studying, different study loads during the semester, and / or different study loads during the days when the practice was scheduled. In addition, near the end of the study, the participants completed a questionnaire which aimed at, among other objectives, identifying any participant who had encountered the target words outside of the class (i.e., in other classes, newspaper, radio, etc.) during the period of the study.

In addition, the number and timing of the stages of the study was carefully designed to conform to the best practice of design of psychological studies examining the spacing effect (see Figure 3.1). In general, the spacing effect is usually studied based on a design that consists of an initial exposure to the target materials followed by only one review, either immediately following or later (e.g., Sobel et al., 2011). In the case of vocabulary in the classroom, that would correspond to a new word occurring for the first time and having relevant information about it explained by the teacher or textbook, followed by one practice (review) exercise done by the students where the word and relevant lexical information about it recurs again.

However, in an actual EFL classroom environment, vocabulary learning can involve introducing the target words to the students then practicing these words on multiple later separate occasions. It is hard to generalise, since in some contexts and with some teachers there may be very little classroom practice targeting the vocabulary that is introduced, which is instead left to students to learn for themselves at home. However, modern textbooks (including associated workbooks and websites) and teaching methods often lead to several follow-up vocabulary-oriented tasks or exercises being available for the new vocabulary that is introduced. These may be implemented by the teacher either all in one lesson (massed) or spread over several lessons and weeks (spaced).

It should be noted, however, that the optimum number of encounters that are required to learn a word to some particular knowledge level, specifically in the context of spaced practice, is not definitively known. However, earlier studies on L2 classroom vocabulary learning suggest that spaced practice leads to better vocabulary learning than massed practice based on three separate practice encounters (e.g., Goossens et al., 2012; Schuetze, 2015). Estimates for the number of separate casual or incidental encounters (see Section 2.5) needed to learn a word are of course much higher (e.g., 12 encounters in extensive reading: Nation, 2014). However, consistent with established massed/spaced research, this study is concerned only with intentional and explicit classroom encounters in vocabulary activities, which promoted the students' knowledge of the target words. Therefore, the words are brought to conscious attention (so noticed) and accompanied by relevant information about them. Consequently, it is reasonable to assume that a minimum of three separate encounters after initial study is advisable for spaced vocabulary learning. In the current study, four review encounters were implemented, so as to allow additionally for practicing words at four strength levels of vocabulary knowledge (i.e., receptive recognition, productive recognition, receptive recall, productive recall) (e.g., Laufer & Goldstein, 2004; Laufer et al., 2004). The aim of this study, unlike previous research, is not just to ascertain how well words are learned in some general unspecified sense, in massed versus spaced learning conditions, but rather how well these words are learned at four defined levels of lexical knowledge, where each repetition systematically involves practice at a different strength level.

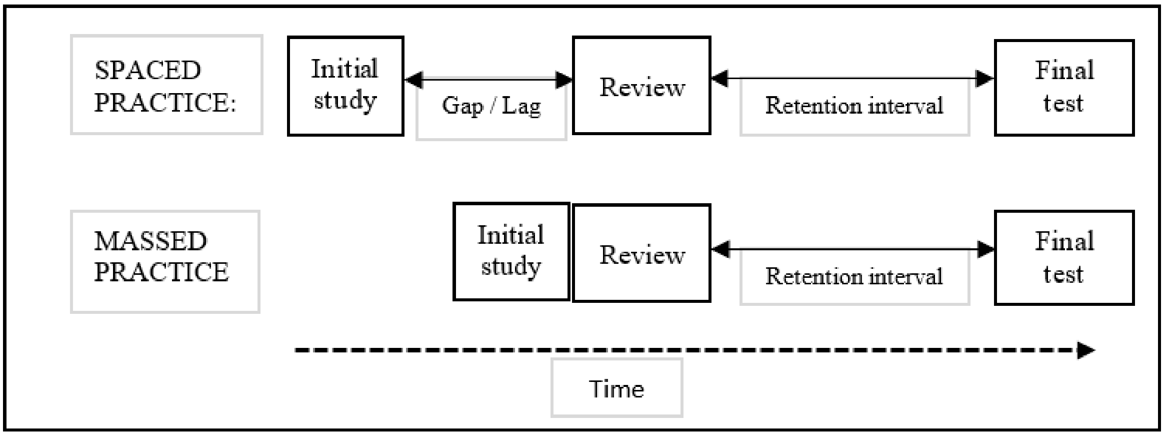


Figure 3.1 The typical experimental procedure for examining the spacing effect based on a single review opportunity. Adapted from Kang (2016, p. 13).

In addition, previous classroom studies often only examined the spacing effect based on delayed post-tests and did not investigate the difference between massed and spaced practice on vocabulary learning through immediate post-tests (i.e., Sobel et al., 2011; Goossens et al., 2012). Therefore, this study examines the spacing effect based on a design that consists of initial study sessions followed by four practice sessions, immediate post-tests, and delayed post-tests (see Figure 3.2).

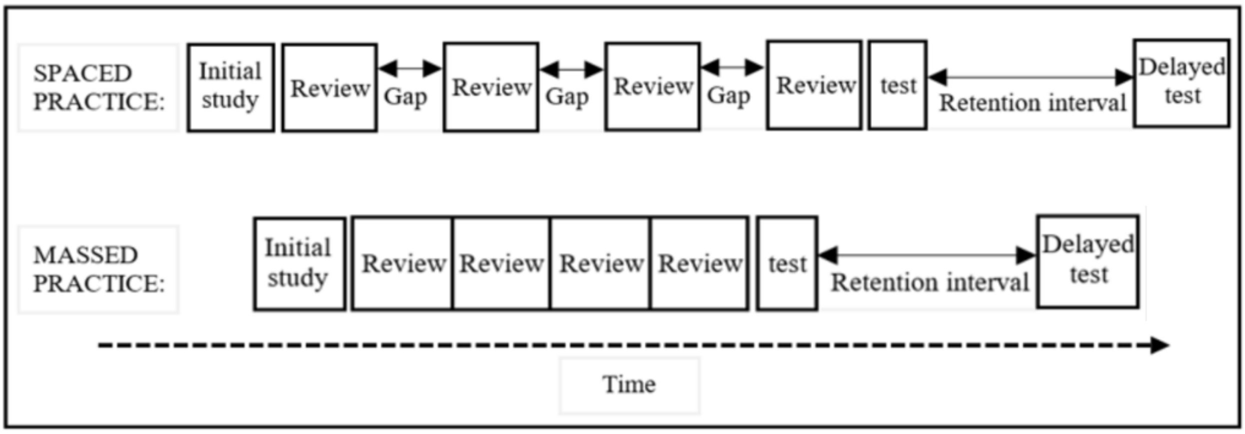


Figure 3.2 Design of the study using four practice tasks, immediate post-tests, and delayed post-tests.

Thirty words were presented and then practiced in four equally distributed learning sessions (spaced practice) and another thirty words were presented and practiced in one learning session (massed practice) (see Section 3.6 for full details of the words). Furthermore, both learning conditions occupied the same overall presentation and practice duration and each class was assigned the same lecture time (i.e., 10am-1pm) each week. Finally, on completion of the four practice tasks in both conditions, immediate post-tests and delayed post-tests were administered. Notably, the retention interval between the last review of the target words and the post-tests and delayed post-tests was the same across the learning conditions, which should ensure that time had an equal effect on retaining words in both learning conditions.

A pilot study was conducted before the main study to inform various aspects of the design of the main study, test the workability of the measurements, and how much time it would take to practice and test the target items. Details of the pilot study are explained in Section 3.9.

3.6 Materials for the main study massed/spaced intervention sessions

3.6.1 Vocabulary selection

The main considerations in the selection of the target words were to choose words which were most likely unknown to the participants (i.e., unlikely to have been encountered inside or outside the classroom, and which have a straightforward translation equivalent in Arabic) (cf. Section 2.6.3). Additionally, the target words should be of similar word frequency and word length (see Sections 2.6.2 and 2.6.4).

These criteria were met through the following steps. First, word selection in terms of frequency level was guided by the estimated vocabulary size of the pilot study participants, who were from the same target population as the main study participants, and by previous research findings on lexical knowledge deficiency among Saudi EFL learners which suggested that at this level, Saudi

EFL students at Taif University have an estimated vocabulary size of 1,408 word families (i.e., around 2,252 lemmas). On this basis, it was deemed that it would be sensible to assume that infrequent words from beyond the 3,000 word-frequency level are unlikely to be known to the target population.

Therefore, as an initial step toward vocabulary selection, a list of the top 5,000 most frequent words from the 450-million-lemmas Corpus of Contemporary American English (COCA) (Davies, 2016) was consulted. A decision was made to use this corpus because it is considered to be the largest, most up-to-date and most genre-balanced native US English corpus (Davies, Wang &, Liu, 2008). Next, in order to control frequency effect and select words that were most likely unknown to the participants, only words from the fifth 1,000 frequency band were extracted.

It was initially intended to examine the spacing effect in relation to all major parts of speech. However, it would be unreasonable to cram a large number of words into a single learning session (for massed practice) while choosing a small word sample from each word class might be insufficient to draw any firm conclusions. Therefore, at the second step, the resulting word list was further examined and words of classes other than nouns and verbs were removed so as to control for word class but still allow a comparison of two word classes.

The third step involved checking the L1 translation (i.e., in Modern Standard Arabic) of each target word. There were two reasons for this step. First, the intended task in the study was to learn a new word form for a meaning the students already knew in L1, rather than to learn a new meaning/concept as well as a form, which would add an extra uncontrolled learning burden. Second, the instrument used in this study tests the participants' strength of knowledge of the target words using translation, which is not satisfactory if there exist no equivalent Arabic words. In order to achieve this step, each target word was checked against two different bilingual dictionaries (i.e., *Oxford Essential Arabic Dictionary*, 2010; *Almawrid English-Arabic Dictionary*, 1995) and words were excluded that have no word of equivalent meaning in Arabic. Only one meaning for each English word was chosen (i.e., the most common one), which had a single word in Arabic that also shared that meaning and, thus, limited the teaching and testing material to one meaning of each word.

At the fourth step, the resulting word list was sorted based on word length (i.e., letter/character count). In general, there are mixed findings about the effect of word length on vocabulary learning, as discussed in 2.6.2. First, different studies used different ways of measuring learning, involving different levels of word knowledge, which could have produced different results depending on what type of measurement was used in each study (e.g., productive knowledge test, receptive knowledge test). Some experts simply say that word length may affect vocabulary learning, as there is “more to remember in long words than in short words” (Nation & Webb, 2011, p. 315), which is surely true more of productive knowledge of the form than other aspects of the word.

Second, depending on the L1 of the learner, the degree of morphological similarity between words in the L1 and L2 (e.g., similar affixes or suffixes), might facilitate learning some longer words where they could be broken down into smaller known or guessable units (Milton, 2009). In the present case, however, the English words were used exclusively in their base, uninflected forms, and there is no similarity between the derivational structure of English and Arabic words. Lastly, the inconsistent findings about word length may also be attributed to the differences in the average length of words and learner’ proficiency levels selected for investigation in the different studies (Al-Masrai and Milton, 2015a).

In general, the contradictory accounts of the effect of word length on vocabulary learning have led some researchers to point out that this issue is anything but straightforward and it might be difficult to isolate the effect of word length from other variables (Laufer, 1997; Milton, 2009). Furthermore, it should be noted that, prior to the pilot study, it was suspected that the target participants might actually know some short infrequent words, at least at the receptive levels of knowledge. Indeed, a preliminary investigation and testing of infrequent words consisting of three to seven letters revealed that short infrequent words (i.e., three- to five-letters words) were indeed more likely to be known at the receptive recall level of knowledge by the target participants than longer ones.

In the light of the above, it was initially intended to include only words that were seven letters long. However, the number of seven-letter verbs in the list was only 26 which fell short of the target goal of 30 words per word class. Therefore, only words that were six letters long (i.e., in their uninflected form) were considered. However, after excluding verbs that were possibly known to the target participants (e.g., *circle, please, credit*, *screen*, *soften*, *repair*), the remaining number of six letter verbs was again below the target goal of 30 verbs. Consequently, in the end, it was decided to include both six and seven letter words in the word list. The final word list consisted of 15 nouns and 15 verbs each six letters long and 15 nouns and 15 verbs each seven letters long.

Finally, the list was randomly split into two lists (i.e., List A and B), each of which consisted of an equal number of words based on word class and length (i.e., 30 nouns and 30 verbs from the 4,001-5,000-frequency level) (see Table 3.1).

It should be noted that the administration of the word lists involved counterbalancing across the instructional conditions. Since the design was repeated measures and not independent groups, the same words could not appear in both learning conditions, yet these words had to be in all respects equivalent words. Gossoons et al. (2012, p. 968) suggested counterbalancing word lists to overcome any possible effect that may arise from certain words being more or less susceptible to spaced versus massed practice. The participants in both classes therefore followed the same instructional procedures in the same order with the only difference being the sequence of the counterbalanced sets of words, which was nested within classrooms. Specifically, list A was used in spaced practice in class one and in massed practice in class two, while list B was used in spaced practice in class two and in massed practice in class one.

List A		List B	
nouns	*verbs*	*nouns*	*verbs*
vaccine	stumble	trauma	reward
texture	provoke	refuge	prevail
glimpse	exploit	harmony	forbid
consent	disturb	dignity	comply
collar	dictate	density	compel
terrain	plunge	pioneer	mutter
helmet	foster	intent	harvest
defeat	donate	dilemma	concede
breeze	condemn	debris	battle
thread	shrink	sponsor	regret
sleeve	persist	permit	isolate
verdict	thrive	vendor	utilize
hostage	insert	mentor	inherit
palace	object	nominee	invade
fatigue	exhaust	escape	descend

Table 3.1 List of target words

3.6.2 Teaching phase activities

Each word was initially presented by the teacher-researcher in spoken and written form along with its L1 equivalent using a paper flashcard. Next, each word was practiced using four types of exercises, which consisted of an L2 definition multiple-choice task (see Figure 3.3), a fill-in-the-gap multiple choice task (see Figure 3.4), an L2-L1 translation task (see Figure 3.5) and an L1-L2 translation task (see Figure 3.6). These exercises were chosen carefully to match the four knowledge strength levels which were to be later tested (see Section 3.3.2). Specifically, exercise one consisted of matching an L2 definition with the target word among multiple-choice alternatives and aimed at practicing the target words on the receptive recognition level. Exercise two is a gap-fill task that consisted of an L2 sentence with the target word among multiple choice alternatives and aimed at practicing the target words on the productive recognition level. Exercise three consisted of supplying an L2-L1 open response translation of a target word in an L2 sentence and aimed at practicing the target words on the receptive recall level. Exercise four consisted of L1-L2 open response translation of an Arabic word in an L1 sentence and aimed at practicing the target words on the productive recall level.

Please choose one word to go with each meaning.			
a strong, hard hat that covers and protects the head			
suburb	helmet	motive	temple

Figure 3.3 Example of exercise one: receptive recognition task

Please fill in the gap with the missing word from the multiple-choice options.			
It is very risky to ride your bike without wearing a ………….			
ritual	insect	combat	helmet

Figure 3.4 Example of exercise two: productive recognition task

Please translate the underlined word	
الجملة	الترجمة
It is very risky to ride your bike without wearing a helmet.	

Figure 3.5 Example of exercise three: receptive recall task

Please write the English translation of the underlined words.	
sentence	translation
لا تنسى ارتداء الخوذة عند قيادة الدراجة.	

Figure 3.6 Example of exercise four: productive recall task

Moreover, the additional words used in the items (distracters and definitions) were crosschecked against BNC-COCA (Nation, 2012) frequency lists using the *VocabProfiler* feature available at Cobb's (2016) Lextutor website[6]. This crosscheck was done in order to minimize use of infrequent words and so ensure comprehension of the content of each exercise, and that the focus of the practice was on the target word and not on other unknown words. For example, the frequency profile of the sentences in exercise two was 84.4% from the 1st 1,000 most frequent words in English, 9.8% from the 2nd 1,000 most frequent words in English and 3.9% from the 3rd 1,000 most frequent words in English.

3.7 Instrumentation

In this section a description is given for each of the instruments used to measure the various different variables involved in this study.

[6] Cobb, T. Range for texts. Accessed 06 Sept 2016 at https://www.lextutor.ca/vp/

3.7.1 Questionnaires

Pencil and paper questionnaires are “any written instrument that presents respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers” (Brown, 2001, p. 6). In general, questionnaires enable researchers to collect factual data (e.g., demographic information, socio-economic status, educational level, and language learning history), behavioural data (e.g., personal history, habits, life-styles) and/or attitudinal data (e.g., attitudes, beliefs, interests, and values) from the respondents.

Depending on the purpose of the questionnaire, a questionnaire can be designed as a closed-ended questionnaire or an open-ended questionnaire. Questionnaires can sometimes include a combination of both types; in such questionnaires open-ended questions usually follow closed-ended questions to enable researchers to gather more information on a topic covered in a closed-ended question (Dörnyei, 2007).

Two questionnaires were used in this study: a biodata and language learning history questionnaire, and a learners’ perception questionnaire. Content validity was assessed based on expert opinions of Mr. Phil Scholfield (personal communication, 2016) who had experience in quantitative research methods such as questionnaires. Consequently, a few changes were implemented relating to questionnaires’ categories and wording.

3.7.1.1 Biodata and language learning history questionnaire

A copy of this questionnaire can be found in Appendix D. The questionnaire was intended to be administered one week before the beginning of the study and includes open-ended questions. Participants were asked to provide their age, how many years they have been learning English, if they stayed in an English-speaking country and for how long, and if they had medical issues that affected their memory. The rationale behind using this questionnaire was to assist in excluding data from participants who might have had extra English learning opportunities, and/or medical issues influencing their memory abilities, so did not belong to the targeted population.

It should be noted that the questionnaire was compiled in English but due to the expected low level of English of the participants, an Arabic translation of the questionnaire was also provided to the participants (Appendix E). Therefore, rather than dictating to the participants which language to use, they could choose to respond in the language they preferred. The translation of the English version of the questionnaire was made by the researcher and a translation expert in the English language department at Taif University.

3.7.1.2 Learners' perception questionnaire

The second questionnaire was intended to measure participants' perceived benefits for learning with, and engagement in using, spaced practice compared with massed practice. A copy of this questionnaire can be found in Appendix B. This questionnaire was intended to be administered at the end of the study (i.e., week 10). It should be noted that it was not possible to adapt existing questionnaires concerned with EFL students' perceptions of spaced learning due to lack of such studies in the classroom vocabulary learning literature. Therefore, a list of statements related to the participants' possible attitudes to using spaced practice and massed practice was created first in English then translated into Arabic. The list was divided into two sets of questionnaire items, one on the participants' perceived benefits of learning with massed and spaced practice, and the other on their perceived benefits of engagement arising from using massed and spaced practice.

The questionnaire consists of eight closed-ended items with an open-ended question at the end. The closed items each took the form of a statement responded to by using a scale from one to five (1: *strongly disagree,* 2: *disagree,* 3: *neutral,* 4: *agree,* 5: *strongly agree*). This format consisting of statements with Likert scale responses is widely used in attitude research (Dörnyei & Taguchi, 2009) and seemed suitable for the purposes of this study.

Additional perceived benefits of using spaced practice or massed practice could be reported by participants in the open-ended item afterwards, where it is hoped to elicit unprompted comments of the students in the form of qualitative data. Similar to the biodata and language learning history questionnaire, an Arabic translation of the questionnaire was provided to the participants (see Appendix C). Additionally, each item in the questionnaire was read aloud and further explanation on how to complete the questionnaire was provided when it was administered.

3.7.2 Vocabulary Size Test (VST) (Nation & Beglar, 2007)

The VST is a multiple-choice receptive vocabulary size test in English (so at receptive recognition strength level). The test was developed to "provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1000 to the 14th 1000-word families of English" (Nation & Beglar, 2007, p. 9). The VST consists of 140 items measuring vocabulary knowledge of written word families in the British National Corpus (BNC) (Nation, 2006), from the first 1,000 to the 14th 1,000 word-frequency level. At each frequency level, ten items were randomly selected and so the estimate of vocabulary size is obtained by multiplying the score obtained by 100.

As previously noted (Section 2.10.2.2), there is a possible tendency for the VST to overestimate vocabulary knowledge (Gyllstad, et al., 2015; Kremmel & Schmitt, 2016). However, despite this possible short-coming of the VST, there are several reasons to use the VST in this study. First, previous studies validating the VST found it to be a highly reliable and valid test of English vocabulary size (i.e., Beglar, 2010; Gyllstad, 2012) provided the VST is not interpreted as if it was an accurate measure of things which it does not claim to measure, such as spoken word knowledge, or ability to recall or produce words. Furthermore, the VST correlates well with general vocabulary proficiency (Gui, 2015).

Second, the VST provides an estimation of vocabulary size knowledge at each word-frequency level from the 1st 1,000 word-frequency level to the 14th 1,000 word-frequency level. The range of word-frequency levels that the VST covers is important because the target items in the study were selected from a lower word-frequency level to minimize the chances of these items being known by the participants (i.e., the 5,000 word-frequency level). Therefore, examining the participants' vocabulary size at the same word-frequency level of the target items, as well as, at lower word-frequency levels would further confirm the expected vocabulary size of the participants and the decision to choose the target items of the study from a low word-frequency level. Third, in terms of classroom time constraints, the test format makes it practical in the target context as it takes a relatively short time to administer.

3.7.3 Design of Vocabulary Achievement Test of Strength (VATS)

The final test in this study was a pen-and-paper Vocabulary Achievement Test of Strength (VATS), with respect to the 60 target words of the study. It was constructed based on a bilingual version of the Computer Adaptive Test of Size and Strength (CATSS) (Laufer & Goldstein, 2004; Laufer, et al., 2004). VATS assesses knowledge of vocabulary meaning at four degrees of strength of knowledge, i.e. productive recall, receptive recall, productive recognition, and receptive recognition, as described in Section 2.10.5.1. Two tests consisting of 120 items were constructed, each for 30 words tested repeatedly in four subtests. Thus, the same words are tested four times.

The reason that two separate tests were required was that the immediate post-test had to be given immediately after the last learning session of each learning condition (i.e., spaced and massed) finished. Since that occurred in different weeks for massed and spaced learning, and on different days of the week for each class (see Table 3.2), it was necessary to make a separate test for the words in list A and list B. Due to the counterbalancing regime which was described earlier, when the spaced learning finished, one class needed to take the test of list A which the class had been studying and the other class of list B. Then next week after the massed learning finished, each class would take the test of the opposite list.

In addition, the immediate and delayed post-tests were constructed to differ in the order of the items and/or distractors in each subtest. This method was done to combat any test memory effect from the immediate test when taking the delayed test. Specifically, using the Research Randomizer online tool[7], the items in all the subtests as well as the distractors in the productive recognition and receptive recognition subtests of the delayed post-tests were placed in a random order different from that used in the immediate post-tests. In total, eight versions of the VATS were created. The formats of the items used in the final versions of the test are as follows.

Productive recall. This part of the test presents the L1 prompt in isolation and requires the participant to supply the target L2 translation equivalent form (see Figure 3.7).

Please supply the English translation for the Arabic word below.

يلقن ……………………………. (answer: dictate)

Figure 3.7 Example of VATS Part 1 (productive recall)

Receptive recall. In this part of the test, the L2 target word form is provided and the participants have to show they know the meaning by translating it into Arabic. (see Figure 3.8).

Please supply the Arabic translation for the English word below.

dictate ……………………………. (Answer: يلقن)

Figure 3.8 Example of VATS Part 2 (receptive recall)

[7] Urbaniak, G. C., & Plous, S. (2016). Research Randomizer v4. 0. 2013 Retrieved from http://www.randomizer.org

Productive recognition. In this part, the L1 translation is provided, representing the target meaning and the participants have to choose the corresponding L2 target word form from the four options. (see Figure 3.9).

Please choose the English translation for each Arabic word. You may choose only one option.

● يلقن

a. concern b. dictate (answer) c. finance d. suppose

Figure 3.9 Example of VATS Part 3 (productive recognition)

Receptive recognition. In the final part of the test, the L2 form is provided and the participants have to choose the corresponding L1 translation from the four options, to show they know the meaning (see Figure 3.10).

Please choose the Arabic translation for each English word. You may choose only one option.

● **dictate**

a.سأم b.أنطلق c.يثبت d.يلقن (answer)

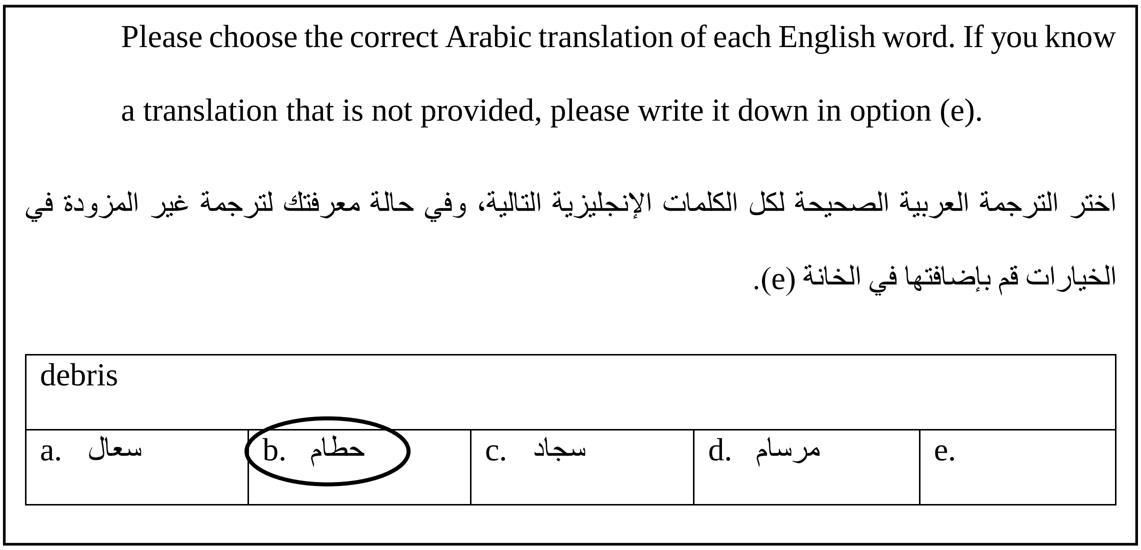
Figure 3.10 Example of VATS Part 4 (receptive recognition)

The selection process for the English distractors in productive recognition items was similar to the selection of the target words. First, all the distractors came from the 5th 1,000 frequency level of COCA, as did the target words, to minimize chances of including known words as distractors. Second, the distractors ranged in length from six to seven letters which corresponded to the letter length of the target words to minimize guessing based on word length. Third, every distractor shared the same word class as that of the key word. Finally, as much as possible, the distractors were assigned to each target word in such a way as to avoid meaning and orthographic similarity. Three distractors were selected for each target word.

The Arabic distractors for the receptive recognition items were selected using an Arabic/English dictionary. Each English target word was assigned three Arabic distractors that shared the word class but differed in meaning from it. Furthermore, I consulted two colleagues who specialize in translation at Taif university. They surveyed a list of the 60 receptive recognition items, each consisting of one target word and three distractors, to spot any issues with the assignment of the distractors. It should be noted that the same selection process for the English distractors in productive recognition items and Arabic distractors in receptive recognition items was followed for the delayed post-test items, with the only difference being re-ordering the items.

3.7.4 Test of previous knowledge of the target words

Additionally, an L2-L1 receptive recognition multiple-choice test was constructed to examine prior knowledge of the target vocabulary. The test consisted of all 60 target words and each item was presented separately along with four Arabic translations (one correct translation and three distractors). The participants were asked to choose the correct answer and/or provide an additional meaning (Figure 3.11). Adding an additional option where the participants provide a different possible translation was to ensure that the target words were not known in any other possible translation that was not included in this pre-test. Using a pre-test at receptive recognition level was deemed suitable to determine if the participants knew the target words as it examines knowledge of the target words at the least difficult strength vocabulary level.



Please choose the correct Arabic translation of each English word. If you know a translation that is not provided, please write it down in option (e).

اختر الترجمة العربية الصحيحة لكل الكلمات الإنجليزية التالية، وفي حالة معرفتك لترجمة غير المزودة في الخيارات قم بإضافتها في الخانة (e).

debris				
a. سعال	[✓ b. حطام]	c. سجاد	d. مرسام	e.

Figure 3.11 Example of L2-L1 multiple-choice pre-test

In order to choose the right translations in Arabic as well as words of the same word class to that of the target word, each target word was checked against two different bilingual dictionaries (i.e., Oxford Essential Arabic Dictionary, 2010; Almawrid English-Arabic Dictionary, 1995). In some cases where more than one Arabic equivalent was found for a target word, the Arabic word that returned the highest number of search hits in Google search engine (2016) was chosen. Next, a list of 60 target words and their Arabic equivalents were checked and approved by a colleague who specializes in translation at Taif university. Furthermore, the Arabic distractors were selected randomly from *Almawrid English-Arabic Dictionary* (1995), provided these distractors shared the same word class as the key word, and possessed similar word length in letters as each key word.

3.8 Procedure

As a lecturer in the English Department at Taif University, I was granted permission to teach a first-semester first-year module (i.e., ‘Reading Skills’) to two intact classes for the duration of three months. This module was chosen because vocabulary teaching was already embedded in the structure of the course. It should be noted that it might be argued to be more appropriate for the study if the normal course focus was unconnected, so no unwanted uncontrolled effect on learning occurred. That, however, would not be pedagogically realistic (i.e., normally in an educational course everything is connected) and would lower the ecological validity of the study (i.e., the findings would not be immediately able to be used to suggest anything about improving any actual classroom course). Nevertheless, steps were taken to ensure that none of the specific English words used in the study (whether targets or distractors) occurred in the normal classes into which the study was embedded.

Accordingly, I followed three steps in an effort to control for possible unwanted uncontrolled occurrence or recycling of the target vocabulary during normal learning sessions. First, during the learning sessions, participants were only exposed to texts from the main course material, which was *Basic Reading Power* (Jeffries & Mikulecky, 2009). This textbook was carefully chosen for two main reasons. The textbook has a unique structure, which features four parts to be used concurrently (i.e., Part 1: Extensive Reading, Part 2: Vocabulary Building, Part 3: Comprehension Skills and Part 4: Thinking Skills). This feature made it feasible to insert the practice exercises of the study into the vocabulary learning sessions and replace the vocabulary building part in the textbook. Furthermore, the textbook specifically targets English learners at a beginning-level (with a 300-word vocabulary), which reduced the chances of exposing participants to any of the infrequent words that were taught in the study.

Second, as much as possible, I carefully prepared and thoroughly reviewed the textbook material before each class to spot any occurrences of the target words in it and substitute any words found with synonyms. In addition to these steps, by the end of the study (i.e., after the delayed post-test), the participants were asked in a survey if they encountered any of the words on the test in the course material, other courses, or outside the class (i.e., on the radio, in a newspaper, etc.).

Third, when learning took place of the study words, participants were not permitted to retain any of the materials used, such as exercise sheets or tests. Nor were they allowed to make any notes in vocabulary notebooks, phones etc. to study later. In this way it was hoped that the effect of any difference between students in autonomous out of class learning strategy ability or motivation, especially between immediate and delayed post-tests, was eliminated.

The study took place in a classroom with rows of desks and instructions were all given in Arabic as well as English. The duration of the study was ten weeks, including the delayed post-tests. In total in both learning conditions, each target word was first presented, then practiced using four different exercises either in massed or spaced condition (for an overview of the procedure, see table 3.2). Both learning conditions occupied the same overall practice time and each class was taught at the same time of day (10am-1pm). One of the two classes used in the study was taught on Tuesdays, the other on Wednesdays. Finally, on completion of the four exercises for any given set of target words, each class sat an immediate post-test and a delayed post-test. The former immediately followed completion of the 4th set of exercises, regardless of whether that was part of a massed or spaced learning session. The latter took place four weeks after the immediate post-test which “should be indicative of learning which is stable and durable” (Schmitt, 2010, p. 157). It should be noted that it is generally unusual for immediate post-tests to follow learning sessions in real-world classrooms. However, it seemed imperative to establish a record of vocabulary achievement at peak attainment which should provide a clearer picture of the effect of both learning conditions on vocabulary gain.

Week	1	2	3	4	5	6	9	10
Spaced practice	no treatment	Presentation Exercise 1	Exercise 2	Exercise 3	Exercise 4	no treatment	no treatment	no treatment
Massed practice		no treatment	no treatment	no treatment	no treatment	Presentation Exercise 1 Exercise 2 Exercise 3 Exercise 4	no treatment	no treatment
Tests	Pre-tests	no tests	no tests	no tests	Post-test (spaced)	Post-test (massed)	Delayed post-test (spaced)	Delayed post-test (massed)

Table 3.2 Procedure of the study

In session one, I introduced myself to the participants in each class and informed them that I would be taking responsibility for the taught course. I also explained the objectives and requirements of the course. Additionally, the participants were informed that they will learn 60 new words and participate in a number of exercises, tests and questionnaires which were for the purpose of this PhD research. The participants were also informed that, while participation in the exercises was mandatory during class hours, they could opt out from the tests and questionnaires at any time without fearing it will reflect negatively on their taught course grades. Furthermore, the students were promised that they would receive ten credit points in the taught course for full participation. Next, the participants were given a participation information sheet and a consent form, which all of them read and signed.

Once all the sheets were read and signed, all forms were collected, and the participants were asked to complete the biodata and language learning history questionnaire (see Section 3.6.1.1). Following that, they were informed that they would shortly take part in a number of tests. The first test was conducted to assess whether they had previous knowledge of the target words (see Section 3.6.4). The results of this test revealed that none of the participants knew the target words. Therefore, no changes were made to the target word list. The second test was the VST (Nation & Beglar, 2007, see section 3.6.2), which measures the participants' receptive vocabulary size. Once the participants completed these tests, all related materials and tests were collected, and formal instruction was resumed.

In weeks two to six when the spaced or massed learning conditions were implemented, the study treatment started 30 minutes after the beginning of the lesson to make sure that all the students were present and to so as be able to transition to the exercises naturally and smoothly during the taught course.

In week two, 30 target words for the spaced practice, with their L1 translations, were first presented to the participants using flashcards (either list A or B, depending on the class). The participants were handed a set of 30 cards each and they followed my directions. First, I read each word and asked the participants to repeat after me. Then, they were instructed to turn the card to the other side and read the L1 translation. This continued until all the 30 words and their L1 translations were covered.

Once the presentation was completed, the flashcards were collected, and participants were handed a multiple-choice exercise (see Figure 2), in which participants were asked to choose the correct L2 target word to match an L2 definition from four options. The task to be performed in the exercise was explained in Arabic, and then completed by the students working individually and without the use of resources such as dictionaries. From session two to session five, the participants practiced the 30 words in the spaced condition (i.e., List A for class one, List B for class two) in the same way as in week two but with a different exercise and without further presentation after week two. Each spaced condition exercise took around 15 minutes to complete and participants were told the correct answers after each exercise to enable them to evaluate their own performance and to ensure that they ended each practice period in possession of the correct information about the target words for retention. Finally, once each exercise was completed and the answer sheets had been collected, in each week the taught course resumed.

In session three, the participants completed an open response L2-L1 translation exercise (see Figure 3) while in session four, they engaged in a fill-in-the-blank exercise in an L2 sentence with the target word offered among multiple choice alternatives (see Figure 4). In session five, the participants completed L1-L2 open response translations in L1 sentences (see Figure 5).

Session five was the final treatment session for the spaced condition and after the completion of the exercise, the participants resumed the taught course for 15 minutes. Then, the immediate post-test for the words learned in the spaced condition was administered. Based on previous piloting of the test, the participants were informed that they had 45 minutes to submit the test papers and leave. The last participant to complete the test took around 40 minutes in class one and 36 minutes in class two.

In session six, as in the spaced condition, participants were presented with 30 new words to be learned, but in the massed condition. Each word was presented and the L1 translation was given exactly as described above for session two. Next, the participants were supplied with an exercise booklet, which consisted of all the target words in the same four types of exercise as described for the spaced condition. Each exercise was handled in the same way as described above, followed by feedback. One hour in all was allowed for the exercises, equivalent to the 15 minutes allowed for each separate exercise in the spaced condition. All exercise booklets and materials were collected once the last exercise was completed. After a 15 minutes break, an immediate post-test for the words learned in the massed condition was administered. Similar to the spaced immediate post-test, the participants were given a maximum of 45 minutes to finish the test.

A delayed post-test was administered for the spaced condition words in session nine (i.e., four weeks after the immediate post-test for the spaced condition words) and another delayed post-test was administered for the massed condition words in session ten, again four weeks after the immediate post-test for those words. The participants received the same tests as in the immediate post-tests but with a different random order of items in each of the four subtests. The delayed post-test was administered during usual classroom sessions following the same procedure as for the immediate post-test.

3.9 Pilot Study and Rationale

The pilot study was conducted between early February and late March 2016 in Saudi Arabia. The reason behind the pilot study was to trial the instruments and treatment schedule in order to reveal and address any issues before conducting the main study.

3.9.1 Pilot participants

The participants of the pilot study were 70 Saudi first year English major students at Taif university in Saudi Arabia, from the same population as that sampled for the main study. The age range of the participants was from 18 to 20 years old, with 62 participants being 19 years old. Most participants began formal English learning in secondary school around the age of 12, which means they had at least six years of previous English study. None of the students had lived and/or studied English abroad. The participants were recruited from two intact classes: class one included 38 participants while class two included 32 participants. However, only data from students who participated in all the sessions of the study were considered, which resulted in a final sample size of 32 participants (i.e., 17 participants from class one and 15 participants from class two). All the participants agreed to take part in the pilot study. The students knew that they were participating in a study, read participation information sheets and signed consent forms.

The pilot study participants in both classes were English language majors. Class one participants were enrolled in a ‘Grammar II’ course while, on the other hand, class two participants were enrolled in a ‘Debate & Discussion’ course.

Based on results of the Vocabulary Levels Test (VLT; Nation, 1990) and Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999), participants knew around 1,064 words receptively and 404 words productively.

3.9.2 Piloting the massed and spaced treatments and their materials

As a lecturer in the English department at Taif university, I received permission to assume responsibility of teaching two intact classes of first year English major students. The pilot study took place in actual classroom settings for the duration of seven weeks, including the delayed post-tests. The participants were informed that, while course attendance and participation was mandatory, they had the right to opt out from including their questionnaire and test data in the pilot study. Additionally, the participants were informed that failure to participate would not reflect negatively on their grades for the actual taught courses, however, they would receive 10 credit points in the taught course for full participation in the study. In addition, the participants were informed that they would take an unspecified test in the final week (i.e., the participants were not informed if it was a test for the study or a test for the normal class content).

In other respects, the pilot materials, and procedures for the two learning conditions and the tests were as described for the main study above except for the following, where the study benefited from the experience of the pilot study to make changes for the main study.

First, the initial presentation of the target words was done using PowerPoint flashcards in the pilot study and using paper cards in the main study. This change was made because, in the pilot study, one of the classes had to be relocated to another classroom because the projector failed to work.

Therefore, in order to avoid similar issues during the main study, paper flashcards were used instead.

Second, the initial presentation of the target words was not done in the same session as any of the exercises in the pilot study. However, in session three (i.e., the first spaced practice session, one week after initial presentation), the participants complained that it was difficult to remember most of the words, which made exercise one very difficult to complete. Therefore, it was decided that in the main study it would be best to introduce the target words and practice them in the same session. Hence in the main study the target words were both presented and practiced in the first session.

Finally, the delayed post-tests in the pilot study were administered one week after the immediate post-tests. This short delay interval was due to practical constraints, as I was only given permission to teach the two classes of students for seven weeks during the pilot study, which did not allow me to increase the delay interval between immediate post-tests and delayed post-test. Clearly a longer period was desirable in order to genuinely test long term retention (Schmitt, 2010) and was implemented in the main study.

3.9.3 Piloting the Vocabulary Achievement Test of Strength

The VATS was administered twice during the pilot study (i.e., immediate post-test, one-week delayed post-test). All participants completed the tests in 45 minutes, except for one participant who took around an hour and 15 minutes to complete the tests. After completing the tests, the participants were encouraged to comment on the tests. The participants provided valuable feedback regarding, for example, page formatting.

In addition, because the participants received the VATS as a whole, it was observed that some of the students tried going back to their answers on lower vocabulary knowledge strength levels (e.g., receptive recognition) in order to answer questions on a higher vocabulary knowledge strength level (e.g., productive recognition). Therefore, a decision was made to not use the VATS as a whole but, instead, in separate parts from the highest strength level to the lowest strength level. This policy was achieved in the main study by first handing-out the productive recall part of the test to all students and the remaining parts on the request of each participant but only after collecting the previous part.

3.9.4 Piloting the Questionnaires

Due to time constraints and availability of participants, only six students responded to the questionnaires. Both questionnaires were administered in the classroom after completing the delayed post-test. All the students chose to complete the Arabic version instead of the English version and it took the participants a maximum of five minutes to complete each questionnaire. No problems arose so no changes were made for the main study (see Section 3.7.1).

3.10 Data Analysis

3.10.1 Analysis of Questionnaire data

As mentioned earlier, this study employed two questionnaires. However, only the learners' perception questionnaire, administered in week ten after the delayed post-tests, was used as a main instrument in the current study. Biodata and language learning history questionnaire, administered prior to week one, mainly assisted only in the selection of the participants of the study and relevant information from it has been reported in Section 3.4 above.

The scores corresponding to the rating scale points used for the closed questionnaire items of the learners' perception questionnaire were inserted into SPSS (Statistical Package for Social Sciences, Version 21.0) for statistical analysis. The responses to the final open question were planned to be submitted to thematic qualitative analysis, so as to identify recurrent themes in what the participants might have answered. However, no responses were received from the participants for this part of the questionnaire.

3.10.2 Internal reliability of the questionnaire

A Factor Analysis and Cronbach's alpha were used to check the internal reliability of the questionnaire. These analyses evaluate whether groups of questions in the questionnaire were related to the same topic in the ways that was planned when the items were chosen. Factor analysis is a statistical procedure “designed to analyse interrelationships within a set of variables or objects” (Reyment & Jvreskog, 1996, p. 71). The Kaiser-Meyer-Olkin (KMO) and Bartlett’s Test measure of sampling adequacy was used first to examine the suitability of Factor Analysis. As can be seen in Table 3.3, the KMO statistic value is between 0.7 and 0.8, which indicates that the data falls into the range of being suitable for Factor Analysis. (Hutcheson and Sofroniou, 1999).

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.758
Bartlett's Test of Sphericity	Approx. Chi-Square	175.4
	df	36
	Sig.	.000

Table 3.3 KMO and Bartlett's Test

Therefore, Factor Analysis should yield informative factors. Furthermore, varimax factor rotation was used to improve the interpretability of the factors (Field, 2017) and the factor analysis resulted in two strong factors with the items loaded onto them as in Table 3.4.

Rotated Component Matrix[a]	Component	
	1	2
D. I recall words better in spaced learning than massed learning.	.828	
A. I memorize words better in spaced learning than massed learning.	.810	
E. I learn words quickly in spaced learning than massed learning.	.759	
B. I memorize more words in spaced learning than massed learning.	.751	
C. I retain words better in spaced learning than massed learning.	.672	
G. I feel more motivated in spaced learning than massed learning.		.833
H. I feel less bored in spaced learning than massed learning		.825
F. I can focus better in spaced learning than massed learning.		.596

Table 3.4 Principal Component Analysis using Varimax with Kaiser Normalization rotation

The grouping of the items matched the grouping of items into two subsets that were planned when the questionnaire was constructed, reflecting respectively perceived benefits for learning and engagement. The following step involved checking the reliability of each factor-defined subset of items. As can be seen in Table 3.5, Cronbach's alphas for the five items in factor 1 and the three items in factor 2 were .84 and .72, respectively, suggesting that the items in each subset have a relativity high internal consistency.

Factor	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
1	.849	.852	5
2	.731	.732	3

Table 3.5 Internal reliability of each questionnaire subscale

3.10.3 Scoring the VST

Scoring the VST was done by following the instructions set out by Nation and Beglar (2007). The VST was scored out of 140 marks by giving one mark for each correct answer and zero for wrong answers. The answer key of the VST was downloaded from Nation's webpage (www.victoria.ac.nz, 2016). Scoring the VST was done by the researcher and another colleague who works as a lecturer in the English department at Taif University. The scores at each frequency level were multiplied by 100 and total score of the VST for each participant were inserted into SPSS for statistical analysis to get the estimated vocabulary size of the participants.

3.10.4 Scoring the VATS

The scoring of the VATS was done by the researcher and another colleague who works as a lecturer in the English department at Taif University. Each of the four strength levels of each version of the VATS were scored out of 30. Each item has only one acceptable answer which had been taught to the participants during the presentation and practice phase. For each answer, the participants received one point for a correct answer and zero for an incorrect answer. In the receptive recognition and productive recognition tests, one correct answer had to be chosen out of four options. In receptive recall, an acceptable response was an L1 translation equivalent of an L2 prompt. In productive recall, an acceptable response was an L2 translation equivalent of an L1 prompt. Responses with spelling errors were considered incorrect answers.

3.10.5 Interscorer and internal reliability of VATS

There are three broad kinds of reliability checks that can be made for any instrument. One is *interscorer reliability* (also known as interrater or interjudge reliability). Interscorer reliability is not often checked for objective test data, like the data in the current study, because scoring is very simple and mechanical, and there is little reason why two scorers should disagree. Indeed, Pearson' correlation (r) between scorers of the immediate post-test was .99 on the receptive recognition and productive recognition tests, and .98 on receptive recall and productive recall tests. Consequently, only one scorer (i.e., the researcher) scored the delayed post-test due to the high interscorer reliability. After thorough revision and double-checking the data, the results were inserted into SPSS for statistical analysis in two forms: scores for each of the two learning conditions and each of the four strength level subtests of the VATS for each participant along with their total score for each subtest; scores for each word class (i.e., noun, verb) in the VATS for each participant at each strength level in each learning condition.

The second type of reliability check is called *test-retest reliability*, which involves giving the same instrument twice to the same participants with a couple of weeks in between to allow them to forget the specifics of the test items and see if they give the same responses again. Time in test-retest reliability checking is expected to pass with the assumption that nothing relevant happens to whatever ability is being measured in the short term. However, the passing of time is one of the experimental variables in the current study and the participant responses, even within a week or two, are expected to change. Therefore, it made no sense to apply this sort of reliability check.

The third type of reliability check is called *internal reliability*, which is used for any instrument that is made of a set of items that are all supposed to be measuring one thing. An internal reliability check clearly applies in the case of the current study. Internal reliability can be assessed with the split half technique (i.e., splitting the set of items into two and seeing how close the scores are between halves). Cronbach's alpha is one of the common statistical methods used to determine internal reliability (Mackey & Gass, 2005), which can be thought of as checking all the possible split halves that could be created (Phil Scholfield, 2016, personal communication). According to DeVellis (2003), an alpha score is very good when it falls between .80 to .90, respectable when it falls between .70 and .80, acceptable when it falls between .65 to .70, and undesirable when it falls between .60 and .65. Any alpha score below .60 is considered unacceptable.

Cronbach's alpha needs to be measured for each condition and group separately, so there were 32 alphas to be calculated, 16 at each time of testing (i.e., immediate post-test and delayed post-test. At each time there were 8 analyses for each list of words: list A massed, at each of 4 levels, list A spaced, at each of 4 levels, list B massed at each of 4 levels, list B spaced, at each of 4 levels.

Test	Time 1: Immediate post-test				Time 2: Delayed post-test			
	Spaced		Massed		Spaced		Massed	
	List A	List B	List A	List B	List A	List B	List A	List B
Productive recall	.544	.556	.270	.497	.100	[2]	.270	.268
Receptive recall	.736	.811	.781	.727	.759	.835	.743	.793
Productive recognition	.714	.722	.818	.749	.708	.766	.750	.700
Receptive recognition	[1]	.609	.605	.672	.533	.658	.779	.615

[1] alpha could not be calculated due to all scores being close to or at 100% correct

[2] alpha could not be calculated due to all scores being close to or at 0% correct

Table 3.6 Reliability of the achievement test items using Cronbach's alpha

As can be seen in Table 3.6, the scores in the tests at the middle two strength levels reached the desirable .7 threshold for reliability. The two extreme levels however exhibited a less consistently strong set of figures. In two instances the scores were too close to all correct or all wrong for alpha to be sensibly calculable, although both those situations in fact indicate high agreement between items. Cronbach's alpha is a correlation-based statistic and does not perform well when there is little or no variation in the responses, because correlation can only exist where there is variation (Tavakol & Dennick, 2011). The low number for participants in relation to the number of items could also have played a role. In reliability checking, ideally the number of participants would exceed the number of items but that was not the case in the current study. A very similar result was obtained by using the split half correlation method on odd and even numbered items (Spearman-Brown, equal lengths) (see Table 3.8).

Test	Time 1: Immediate post-test				Time 2: Delayed post-test			
	Spaced		Massed		Spaced		Massed	
	List A	List B	List A	List B	List A	List B	List A	List B
Productive recall	.538	.567	.358	.476	.107	[2]	.358	.325
Receptive recall	.742	.862	.802	.765	.763	.853	.801	.799
Productive recognition	.769	.712	.709	.764	.754	.768	.699	.745
Receptive recognition	[1]	.817	. 754	.616	.800	.743	.630	. 607

[1] alpha could not be calculated due to all scores being close to or at 100% correct
[2] alpha could not be calculated due to all scores being close to or at 0% correct

Table 3.7 Reliability of the achievement test items by the split half method (odd-even items) using the Spearman-Brown coefficient

Importantly, productive recall exhibited the lowest reliabilities, including four alphas below .3. This low reliability, however, could be due to the fact that productive recall was the hardest strength condition, which in turn led to the participants getting large numbers of items wrong and, as stated above, made correlations poor (Tavakol & Dennick, 2011). Furthermore, the participants may have responded inconsistently on the few items that they got right. Since the same words occurred in all four strength levels, any differences in reliability at different strength levels are presumably due to the strength level of the items in relation to learner proficiency rather than the choice of words. There is little that can be done to such participant-based unreliability, since it is not customary to reject participants who respond unreliably in one or two experimental conditions (Phil Scholfield, 2016, personal communication).

In some of the low alpha instances, omission of one or two items would have raised the alpha slightly, but not dramatically up to a level such as .7, so that avenue was not pursued. Furthermore, in the interests of parallelism of the research design, I did not wish to omit scores for some items in some conditions but retain those items in other conditions.

It is notable incidentally that the patches of unreliability in Tables 3.6 and 3.7 are not confined to massed or spaced conditions, nor to one time or indeed one list of words. It was not possible to find reliability statistics in comparable other studies as such statistics are rarely reported in journal articles. Laufer et al., (2004), however, did report such statistics. Interestingly, Laufer et al. (2004) found patches of low reliability in the two less demanding recognition tests of CATSS. According to Laufer et al. (2004), the reason behind the low reliabilities was due to similarity of the proficiency levels of the participants, which could have generated lack of variation in the data and reduced apparent reliability.

In conclusion, the tests should be accepted as predominantly moderately reliable since 76% of the alphas were above .6 at least, and the areas of unreliability seemed to involve factors that were unavoidable due to the participants being quite a homogeneous group in terms of their level of vocabulary attainment, as “the lack of variance generally results in low internal reliability values” (Brown, 1983: 86).

3.10.6 Statistical analysis

Only data from students who took part in all learning sessions and test occasions were included in the analyses, which resulted in a final sample size of 49 participants (i.e., 26 participants from class 1 and 23 participants from class 2). First, the Kolmogorov–Smirnov one sample test with Lilliefors correction as well as the Shapiro-Wilk test were used to check the normality of the data, i.e., to test whether the normality assumption of ANOVA was met. The normality checks indicated that none of the sets of scores for the various tests used in the study in any condition exhibited the normal distribution shape assumed by GLM/ANOVA. Due to lack of normality, therefore, it was decided to use the ordinal option within the Generalized linear model (GZLM) for inferential tests, which does not require normality. Particularly, the 'Generalized Estimating Equations' (GEE) option within GZLM is equivalent to the Repeated Measures option within GLM/ANOVA and outputs significance levels for main effects and interactions which can be interpreted in an exactly parallel way to reports of repeated measures ANOVA. Instead of an *F* statistic, however, it employs the Wald chi squared. Furthermore, the ordinal option within GZLM is similar to well-known nonparametric choices like the Wilcoxon or Friedman test in that it treats the scores as rank ordered, thus avoiding any assumptions about the shape of the population distributions.

Since the questionnaire data was based on rating scale responses which are essentially ordinal rather than interval in nature, non-parametric statistics were also used to analyse the questionnaire data. Furthermore, in analyses where more than two values of a variable needed to be compared pairwise, particularly the four strength levels, Bonferroni adjusted significances are reported. Due to performing multiple post hoc paired comparisons on the same data, the Bonferroni adjustment was used to protect against over-claiming significant results.

3.11 Summary

This chapter provided a description of the study detailing the participants, instruments, procedure, and data analysis. It also covered how the instruments and treatment were piloted. Outcomes from the pilot study were implemented in the main study. The following chapter will provide details of the results of the main study.

Chapter Four - Results

4.1 Introduction

This chapter reports the findings of the main study. Since none of the target words were known when tested before the intervention, the data from the immediate post-tests just after the end of each of the two practice phases (i.e., spaced and massed) was used to measure initial learning of word knowledge at four lexical strength levels (i.e., receptive recognition, productive recognition, receptive recall, and productive recall). Retention after four weeks under each treatment could then be ascertained from the same participants via their delayed post-test data. Attitudes to the two vocabulary practice treatments were obtained from responses to the learners' perception questionnaire. In order to present the results as directly and as clearly as possible, the results are organized according to the order of the research questions (Section 2.12).

4.2 RQ 1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?

The descriptive statistics of the immediate post-tests scores are reported in Table 4.1. It is apparent from Table 4.1 that scores for spaced items were higher than scores for massed items at all four strength levels of vocabulary knowledge. The differences in mean scores were greater than 3 words at the two higher strength levels (unsupported recall), while the two low strength levels (recognition with multiple-choice support) had a difference of less than 1.5 words. It is worth pointing out that the small SDs and very high mean scores in the receptive recognition test in both learning conditions indicate a ceiling effect. If the median scores were considered, again the differences were greater than 3 words at the higher strength levels, while the productive recognition strength level had a difference of only one word. There was no difference, however, between the median score of massed and spaced practice at the receptive recognition level. Overall then, the effect of learning condition in terms of difference in number of words learnt was at least twice as much at the more demanding strength levels than at the less demanding strength levels.

		Strength test			
		Productive recall	*Receptive recall*	*Productive recognition*	*Receptive recognition*
Massed	Mean	3.08	11.43	26.67	29.29
	Median	3.00	10.00	28.00	30.00
	Std. Deviation	2.07	5.30	3.24	1.31
	Skewness	0.14	0.28	-1.18	-2.31
	Std. Error of Skewness	0.34	0.34	0.34	0.34
	Kurtosis	-0.76	-1.31	0.87	5.58
	Std. Error of Kurtosis	0.67	0.67	0.67	0.67
	Minimum	0	4	17	24
	Maximum	8	21	30	30
Spaced	Mean	6.39	15.76	27.96	29.69
	Median	6.00	16.00	29.00	30.00
	Std. Deviation	2.78	5.32	2.52	0.65
	Skewness	0.25	-0.17	-1.42	-2.42
	Std. Error of Skewness	0.34	0.34	0.34	0.34
	Kurtosis	-0.30	-0.91	1.25	6.12
	Std. Error of Kurtosis	0.67	0.67	0.67	0.67
	Minimum	0	6	21	27
	Maximum	13	26	30	30

Table 4.1 Descriptive statistics for immediate tests scores on all strength levels

In a normal distribution of data, the values of skew and kurtosis are zero and any value above or below zero generally indicates that the distribution of the data deviates from normal (Fields, 2018). In Table 4.1, skewness values for scores in both learning conditions seem to be positive at the two higher strength levels (indicating many low scores in the distribution) and negative at the two lower strength levels (indicating many high scores in the distribution). Furthermore, kurtosis values are strongly positive at the lowest strength level, in both learning conditions. This indicates that the distribution of scores is pointier than normal (i.e., there are few scores in the tails, again showing ceiling effect).

An overall GZLM-GEE analysis was conducted with type of practice (i.e., massed vs spaced) and strength of vocabulary knowledge tested (i.e., receptive recognition, productive recognition, receptive recall, productive recall) as the independent variables, and the immediate post-test scores as the dependent variable (see Table 4.2). The result indicated that both factors had a highly significant main effect. Furthermore, the interaction effect was also significant which indicated that the differences between strength scores were not the same in each learning condition.

Source	Wald Chi-Square	df	Sig.
Learning condition	36.67	1	<.001
Strength	270.76	3	<.001
Learning condition * Strength	17.76	3	<.001

Table 4.2 GZLM analysis of Immediate Post-test scores

A graph was generated displaying the mean scores at all strength levels in both massed and spaced conditions (Figure 4.1) in order to further understand the interaction effect. It can be observed from the graph that the spaced practice scores were higher than the massed practice scores at each strength level and strength scores rose in the way predicted by Laufer and Goldstein (2004) across the four different levels of strength.

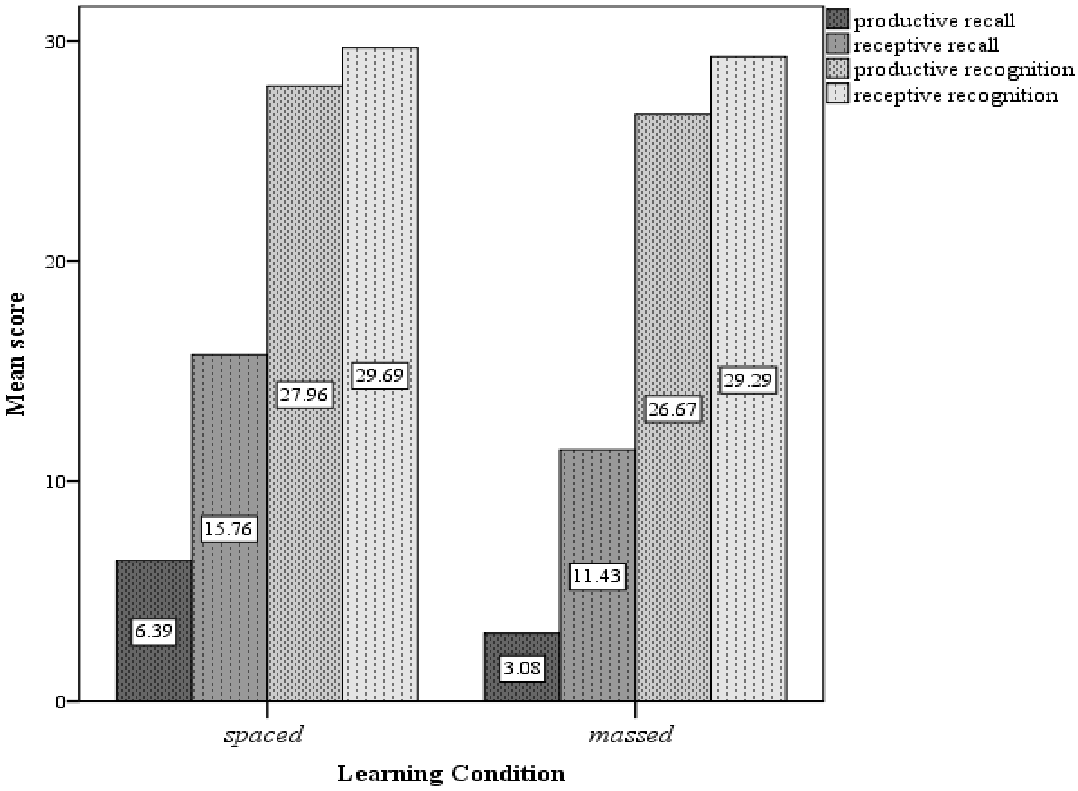


Figure 4.1 Mean vocabulary scores at each strength level for Massed and Spaced conditions: Immediate post-tests.

The interaction effect was significant due to the massed-spaced differences being greater for recall (especially receptive recall) than for recognition (especially receptive recognition). The mean difference between practice types was 3.3 words for productive recall, and 4.4 for receptive recall, but only 1.3 for productive recognition and 0.4 for receptive recognition. This suggests that the impact of spaced learning is greater (initially, at any rate) where the task is made harder by being open response than where it is easier because alternatives are provided, one of which is correct. This result may also be attributed to ceiling effect, since mean scores for recognition were very close to the maximum possible score of 30 (100% correct). In particular, mean scores in the receptive recognition test were 29.69 in the spaced condition and 29.29 in the massed condition, which is very high. Since at these strength levels many students initially learned all the target words, regardless of practice condition, there was only a very small part of the score scale within which a massed-spaced difference could appear for the recognition subtests.

In order to further establish whether significant differences between learning conditions existed at all four strength levels, Wilcoxon Signed-Rank Tests (with Bonferroni adjustment) were conducted to compare vocabulary test scores between the two practice conditions (see Table 4.3).

	productive recall (massed) - productive recall (spaced)	receptive recall (massed) - receptive recall (spaced)	productive recognition (massed) - productive recognition (spaced)	receptive recognition (massed) - receptive recognition (spaced)
Z	-5.603	-4.422	-3.335	-2.486
Asymp. Sig. (2-tailed)	<.004	<.004	<.004	.052

Table 4.3 Comparisons between massed and spaced conditions at each strength level: Immediate Post-test

Results indicated that the spaced scores were in fact significantly higher than the massed scores at only three strength levels (i.e., productive recall, receptive recall, and productive recognition) and fell just short of significance at the easiest level, receptive recognition (*p* = 0.052), where ceiling effect is particularly strong.

In addition, Table 4.3 showed that the Wilcoxon Z statistics decreased successively across the four strength levels from productive recall to receptive recognition. This result partially matched the decrease in the mean difference between the two learning conditions across the strength levels and illustrates again the interaction effect noted above.

4.3 RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test?

The descriptive statistics (means and standard deviations) of the delayed post-test scores are reported in Table 4.4. Regarding the distribution of data, the skewness value for the scores in the spaced condition at receptive recall level is close to zero, which indicates normality. However, skewness values for the scores in the massed condition at the two higher strength levels and in the spaced condition at the highest strength level are positive (i.e., many low scores in the distribution), while scores at the two lower strength levels are negative (i.e., many high scores in the distribution). Furthermore, the kurtosis values at the three higher strength levels in the massed condition and at the receptive recall level in the spaced condition are negative, which indicate that the distribution of the scores is somewhat flatter than normal (i.e., more scores in the tails). The positive values of kurtosis most notably at the receptive recognition level in the spaced condition indicate that the distribution is pointier than normal (i.e., fewer scores in the tails).

		Strength test			
		Productive recall	*Receptive recall*	*Productive recognition*	*Receptive recognition*
Massed	Mean	.90	6.84	24.22	27.16
	Median	1.00	6.00	25.00	28.00
	Std. Deviation	.96	4.54	3.95	2.72
	Skewness	.65	.38	-.80	-1.26
	Std. Error of Skewness	.34	.34	.34	.34
	Kurtosis	-.75	-.94	-.27	1.08
	Std. Error of Kurtosis	.67	.67	.67	.67
	Minimum	0	0	14	19
	Maximum	3	17	30	30
Spaced	Mean	4.16	14.45	27.65	29.16
	Median	4.00	15.00	29.00	30.00
	Std. Deviation	2.26	5.53	2.55	1.30
	Skewness	.60	-.01	-1.31	-1.75
	Std. Error of Skewness	.34	.34	.34	.34
	Kurtosis	.55	-.84	.75	2.75
	Std. Error of Kurtosis	.67	.67	.67	.67
	Minimum	0	5	21	25
	Maximum	10	26	30	30

Table 4.4 Descriptive statistics for delayed tests scores on all strength levels.

An overall GZLM-GEE analysis was conducted with type of practice (i.e., massed vs spaced) and strength of vocabulary knowledge tested (i.e., receptive recognition, productive recognition, receptive recall, productive recall) as the independent variables, and the delayed post-test scores as the dependent variable (see Table 4.5).

Source	Wald Chi-Square	df	Sig.
Learning condition	99.715	1	<.001
Strength	255.754	3	<.001
Learning condition * Strength	8.797	3	.032

Table 4.5 GZLM analysis of Delayed Post-test scores

The results indicated that learning condition and strength factors again had highly significant main effects. The interaction effect was again significant, but the size of this effect was considerably smaller than that on the immediate post-test. Furthermore, judging from the Wald Chi-Square figures, the effect of strength was similar to that in the immediate post-test, while the effect of learning condition was much stronger in the delayed post-test than in the immediate post-test.

A graph was generated displaying the mean scores at all strength levels in both massed and spaced conditions to further understand the interaction effect (Figure 4.2). This graph suggested that the interaction effect again was significant due to the difference between spaced and massed practice scores not always being the same at different strength levels. In this instance, however, they are not systematically greater at both recall strength levels than at both recognition strength levels. The mean differences between practice conditions were 3.3 for productive recall, 7.6 for receptive recall, 3.5 for productive recognition and 2.0 for receptive recognition. Thus, receptive recall level showed the greatest advantage for spaced learning over massed learning: impressively on average twice as many words were correctly remembered at this level of knowledge after original spaced learning than through original massed learning. The other three strength levels however showed similar much smaller differences between scores for words learned in the different conditions at the end of the study, compared with a zero baseline at the start.

Once again there could be some ceiling effect at work here accounting for the smaller massed-spaced differences between recognition scores, although the recognition scores were not as close to the possible maximum of 30 as those in the immediate post-tests. There could also be floor effect accounting for the lack of great difference between massed and spaced productive recall scores. Where scores generally are close to the lower limit of the score scale on any measure, just as when they are close to the upper limit, it is well known that differences between conditions tend to become small due to lack of available space on the scale for them to become apparent.

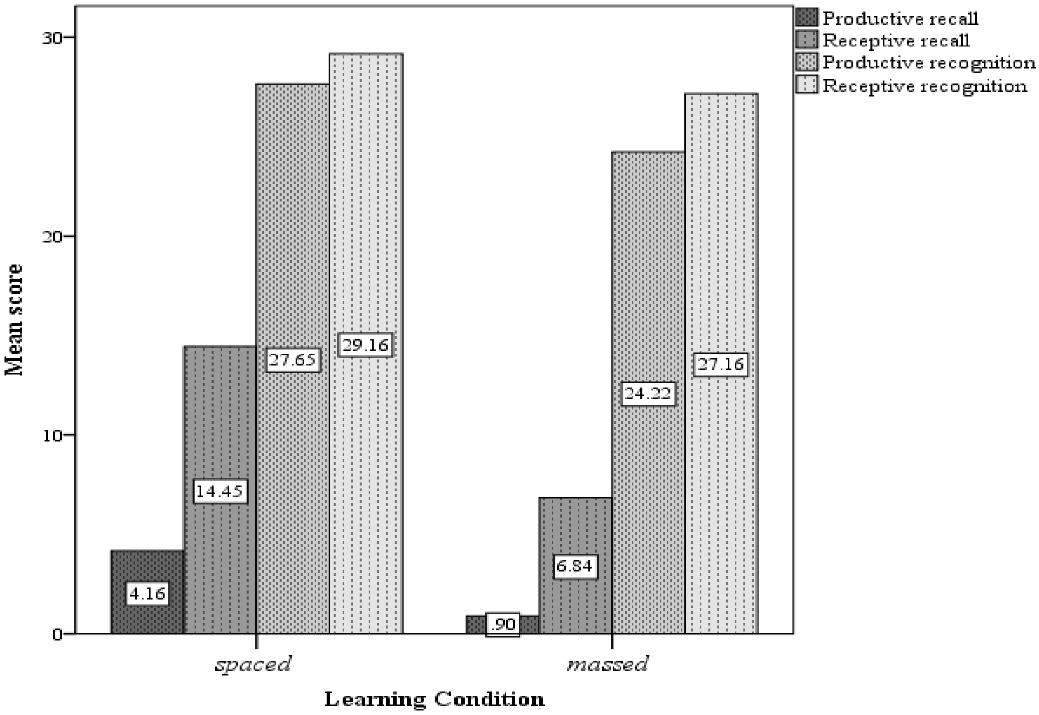


Figure 4.2 Mean vocabulary scores at each strength level for Massed and Spaced conditions: delayed post-tests.

Furthermore, Wilcoxon Signed-Rank Tests were again conducted comparing learning conditions at each strength level separately to establish whether significant differences between learning conditions existed at all four strength levels (see Table 4.6). Results suggested that the two learning conditions differed significantly at every strength level. Specifically, the spaced scores were significantly higher than the massed scores at each strength level.

	productive recall (massed) vs productive recall (spaced)	receptive recall (massed) vs receptive recall (spaced)	productive recognition (massed) vs productive recognition (spaced)	receptive recognition (massed) vs receptive recognition (spaced)
Z	-5.998	-5.866	-5.310	-4.794
Asymp. Sig. (2-tailed)	<.004	<.004	<.004	<.004

Table 4.6 Comparisons between Massed and Spaced conditions at each strength level: Delayed Post-test.

The Wilcoxon Z statistics decreased again successively across the four strength levels from the most demanding to the least demanding. This again did not fully match the decrease in mean difference between the two learning conditions across the strength levels. As mentioned above, it was no longer the case that the differences were greater at the two higher strength levels than at the two lower strength levels. In fact, the most demanding level (i.e., productive recall) exhibited a difference of only 3.3 words between learning conditions, which was the same difference as was seen in the immediate post-test and similar to the difference for delayed productive recognition at 3.5 words.

4.4 RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength level between the immediate post-test and the delayed post-test?

This research question concerns the retention, or lack of it (forgetting/attrition), of what was initially learned in the learning conditions, rather than how much was learned between the baseline (= zero knowledge in this study) and the end of the initial learning or delayed periods (RQs 1 and 2 above). Hence, the changes in scores between the immediate and delayed tests are examined. First, a line graph was generated displaying the mean scores at all strength levels in both massed and spaced conditions, and on both the immediate and delayed post-test (Figure 4.3). This line graph will be used to help interpret the results for this research question.

From the downward slope of all the lines in Figure 4.3, it can be seen at once that, as very widely reported in studies of learning of anything, there was at all strength levels, and regardless of initial learning condition, some loss of knowledge (forgetting) between the immediate and delayed tests. Particularly, scores for the massed condition seem to have fallen more than scores in the spaced condition. In order to obtain the fullest picture of the effects of the two learning conditions on learning, a three-way GZLM analysis was performed, with three repeated measures factors: immediate versus delayed post-tests, the four strength levels of lexical knowledge, and the two learning conditions (see Table 4.7).

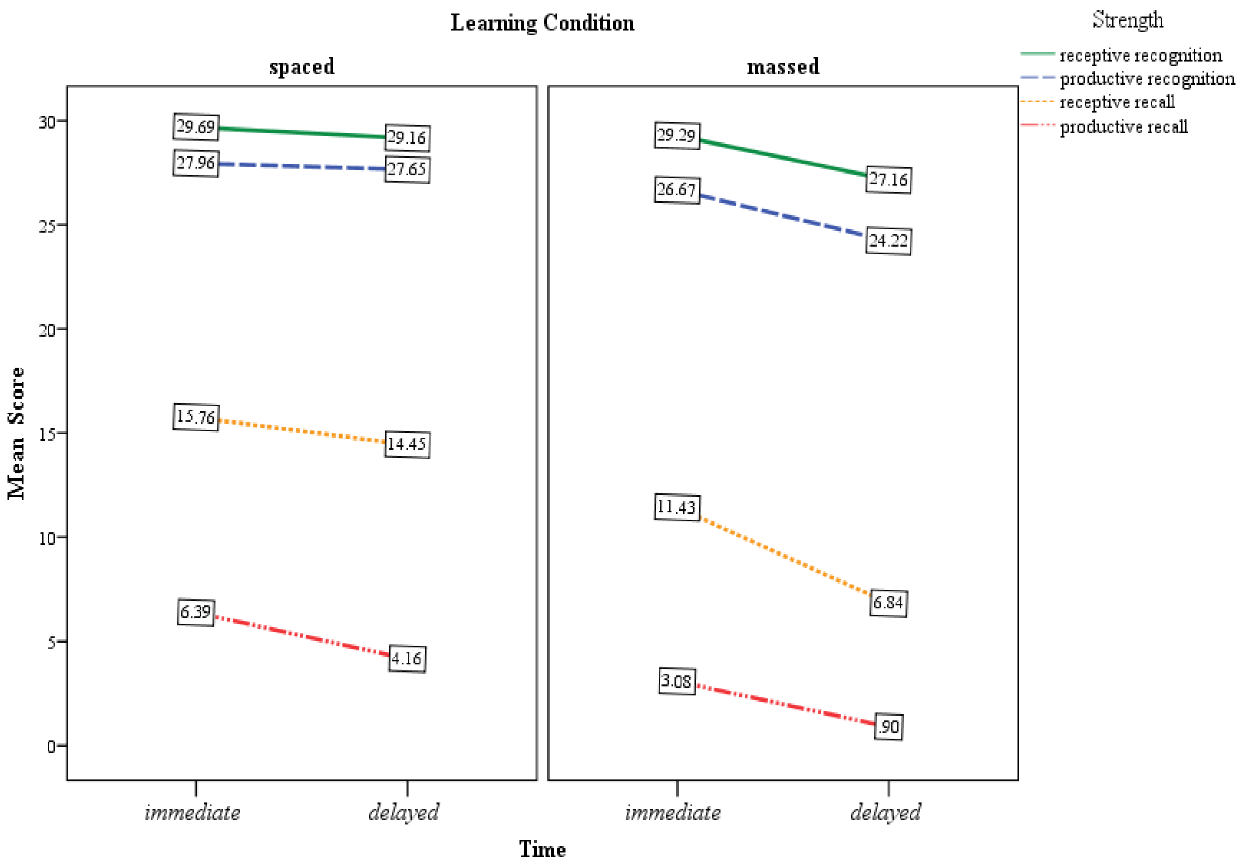


Figure 4.3 Mean scores of immediate and delayed post-tests by learning condition and strength level (mean scores MAX=30).

The overall GZLM analysis using the ordinal option indicated that there was indeed a significant overall decline in scores between the immediate post-test and delayed post-test regardless of learning condition and strength level (see Figure 4.3). Furthermore, there was a significant difference between massed and spaced practice regardless of time and strength, with higher retention scores for the spaced condition, as would be anticipated from the earlier analyses. There was also a significant difference between strength levels regardless of learning condition and time.

Source	Wald Chi-Square	df	Sig.
Learning condition	71.311	1	<.001
Time	158.454	1	<.001
Strength	282.940	3	<.001
Learning condition* Time	59.178	1	<.001
Learning condition * Strength	13.924	3	.003
Time * Strength	19.016	3	<.001
Learning condition * Time * Strength	6.340	3	.096

Table 4.7 Overall GZLM analysis of immediate and delayed post-test scores, by learning condition and strength level

Moving to the interaction effects, all two-way interaction effects were significant, while the three-way interaction effect did not achieve significance. The interactive effect of learning condition and strength on scores has, in effect, already been reported in answering RQ1 and RQ2. What primarily needs attention to answer RQ3 is therefore the learning condition by time effect which is also in fact by far the strongest of all the interaction effects (Wald chi sq. = 59.18), and the strength by time effect.

The learning condition by time effect indicated that retention scores between the immediate and delayed post-tests were not the same in both learning conditions, regardless of strength. The general nature of this can be seen from Figure 4.3 and in a different way Figure 4.4. For all learning strengths except the most demanding (productive recall), the fall in scores between immediate and delayed post-tests was greater for words learned in the massed condition compared with the spaced condition. This fall in scores led to the overall finding that, for all strengths taken together, forgetting was significantly greater of words originally learned in the massed condition than of those learned in the spaced condition. The three-way interaction effect was not statistically significant, which suggests that the learning condition by time effect did not in fact differ significantly among strength levels.

The general pattern of forgetting/attrition over time after the two learning conditions may be seen again in Figure 4.4. Clearly, receptive recall exhibited the greatest difference in forgetting between the two learning conditions. On average, around four words were forgotten at this level of knowledge after massed learning, but only about one after spaced learning, during the immediate-delayed post-test period. By contrast, at the level of productive recall the average loss was about two words regardless of learning condition.

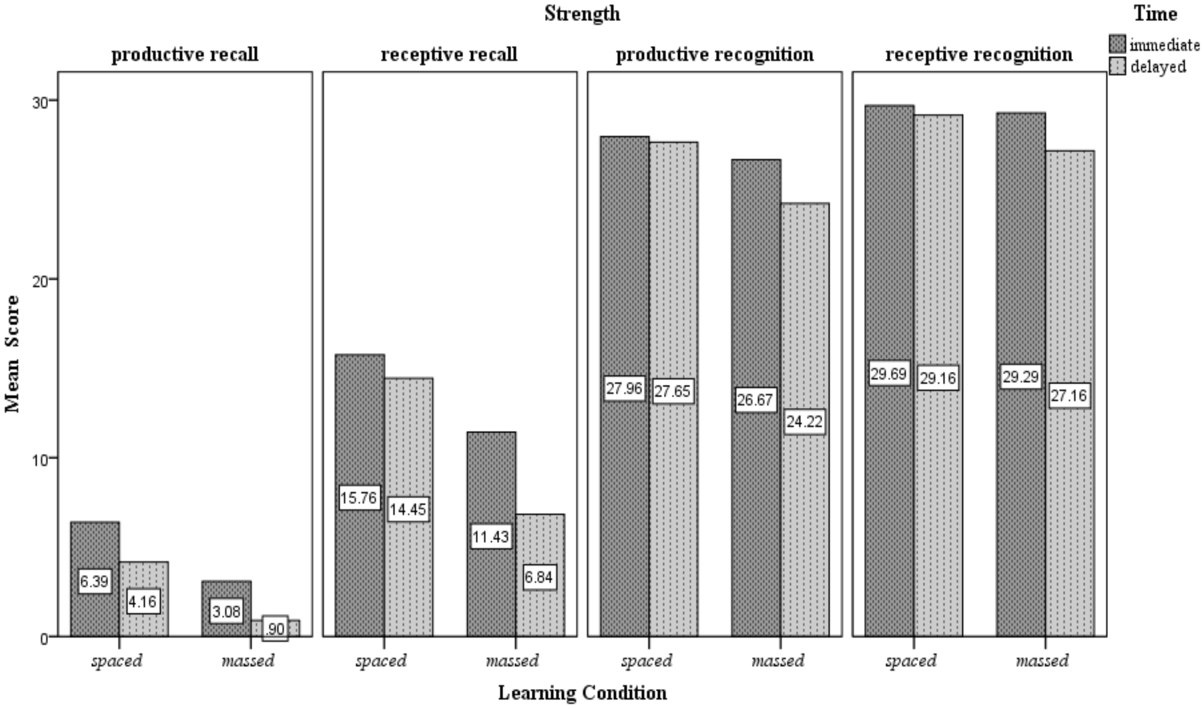


Figure 4.4 Mean massed and spaced vocabulary scores at each strength level, and on the immediate and delayed post-test

It may be concluded, then, that massed practice not only produced lower scores than spaced practice on the immediate post-tests, but also yielded a greater fall in scores than spaced practice did between immediate and delayed post-tests. At the hardest strength level, differences between learning conditions were less easy to appear which was perhaps because scores were generally quite low (i.e., floor effect). Furthermore, it was noticeable descriptively that the difference in vocabulary loss was greatest at the receptive recall strength level, which was the level where scores were more often obtained that were distant from the extremes of the scale (floor = zero, ceiling = 30).

Given that it has emerged from the analyses used above to answer RQs 1-3 that there were some interesting differences between the four strength levels in both main and interactive effects, it was decided to explore these differences further by decomposing strength into two binary variables or dimensions rather than treating it simply as one 4-valued variable. As discussed in section 2.10.3.1, the notion of strength of knowledge that is being used in this study makes use of two oppositions to create four levels of strength. First there is the distinction between receptive knowledge, where an L2 word form is seen and the meaning has to be retrieved by the learner, and productive knowledge, where a meaning is provided (e.g., as an L1 word or an L2 paraphrase/definition) and the L2 word form has to be retrieved, which is usually found to be harder. That I will refer to as the ‘receptive-productive dimension’. Second, there is the distinction between two levels of independence of the learner’s mastery of the knowledge underlying the retrieval: either the learner can retrieve the word with no help provided where the meaning or form has to be retrieved unaided from memory (recall), or alternatives are provided, one of which is correct, and a choice simply has to be made between them (recognition). That will be referred to as the ‘recall-recognition dimension’. Consequently, it might be possible now to consider which of those oppositions plays the greater role in determining scores that learners obtain.

Means for the strength dimensions are displayed in Table 4.8. It may be seen at once that the two dimensions differ considerably. The receptive-productive dimension yielded differences between productive and receptive knowledge of around 5 words on average, regardless of practice type or whether the immediate or delayed test is considered. By contrast the recall-recognition difference has mean scores for recall and recognition nearer to the ends of the score scale, especially for recognition at the top end, and differing by around 19 words.

Strength dimensions		Time			Learning condition		
		Immediate	Delayed	Diff.	Spaced	Massed	Diff.
1	productive	16.03	14.23	1.8	16.54	13.72	2.82
	receptive	21.54	19.4	2.14	22.27	18.68	3.59
2	recall	9.16	6.59	2.57	10.19	5.56	4.63
	recognition	28.4	27.05	1.35	28.62	26.84	1.78

Table 4.8 Means for separate strength dimensions by time and learning condition

This suggests that the impact on scores of the recall-recognition dimension exceeds that of receptive-productive dimension and indeed the GZLM analysis for the main effects of Strength type (see Table 4.9) supports that: the Wald statistic for the recall-recognition dimension is 25% greater than that for receptive-productive dimension. In short, being given multiple choice questions (vs open response questions) helps the student achieve higher scores far more than being asked for receptive information rather than productive, regardless of learning mode and time.

Source	Wald Chi-Square	df	Sig.
Learning condition	71.311	1	<.001
Time	158.454	1	<.001
Receptive-productive	207.819	1	<.001
Recall-recognition	258.041	1	<.001
Learning condition * Time	59.178	1	<.001
Learning condition * Receptive-productive	2.663	1	.103
Learning condition * Recall-recognition	12.841	1	<.001
Time * Receptive-productive	0.256	1	.613
Time * Recall-recognition	4.554	1	.033
Receptive-productive * Recall-recognition	63.789	1	<.001
Learning condition * Time * Receptive-productive	2.434	1	.119
Learning condition * Time * Recall-recognition	0.036	1	.850
Learning condition * Receptive-productive * Recall-recognition	0.338	1	.561
Time * Receptive-productive * Recall-recognition	16.575	1	<.001
Learning condition * Time * Receptive-productive * Recall-recognition	0.331	1	.565

Table 4.9 Overall GZLM analysis of immediate and delayed post-test scores, by learning condition and two strength dimensions

As may be seen in more detail from Table 4.9, both dimensions of knowledge then have an overall significant main effect on scores, with the receptive-productive dimension yielding generally lower Wald statistics than the recall-recognition dimension. The interaction between the two dimensions also has a highly significant effect on scores: the effect of one dimension is not the same when combined with each value of the other. Rather, as seen earlier, the receptive-productive distinction makes only a small difference in scores where recognition is tested while it makes a larger difference where recall is required.

Of special interest, however, are the interaction effects where the two strength dimensions differ. These are the interactions with learning condition and with time, and in both cases the recall-recognition dimension yields a significant interaction effect while the receptive productive distinction does not. What this suggests is that the receptive-productive dimension has an effect that is the same regardless of the learning condition or whether scores come from the immediate or delayed test. The recall-recognition dimension, however, is sensitive to both those variables and there are different differences between recall and recognition under the different learning conditions and on the two test occasions. What emerges is that the difference between recognition and recall scores is not the same for items learned spaced and massed, nor in both the immediate and the delayed tests, while the comparable differences between production and reception scores are not significant (see Table 4.9). As Table 4.8 shows, the recognition-recall differences in scores are greater in the massed than the spaced condition (respectively 21.28 and 18.43) and greater in the delayed post-test than the immediate one (respectively 20.46 and 19.24). Furthermore, the recognition scores differ little between learning conditions or times, while recall scores differ slightly more. This could, however, as already discussed, be due to ceiling effect.

4.5 RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed?

The descriptive statistics (means and standard deviations) of the immediate and delayed post-tests scores by part of speech/word class are reported in Table 4.10 and Table 4.11.

		Massed				Spaced			
		productive recall	productive recall	productive recognition	receptive recognition	productive recall	productive recall	productive recognition	receptive recognition
nouns	mean	2.35	6.55	13.55	14.84	4.04	9.20	14.41	14.86
	median	2.00	7.00	14.00	15.00	4.00	10.00	15.00	15.00
	SD	1.30	2.74	1.92	0.47	2.03	3.16	0.86	0.35
	min	0	1	7	13	0	4	12	14
	max	6	14	15	15	8	14	15	15
verbs	mean	1.14	4.63	12.92	14.53	1.94	6.80	13.76	14.76
	median	1.00	4.00	14.00	15.00	2.00	7.00	14.00	15.00
	SD	1.21	3.11	2.14	0.96	1.49	3.03	1.41	0.60
	min	0	0	6	10	0	2	10	12
	max	4	12	15	15	5	13	15	15

Table 4.10 Descriptive statistics for word class scores in the immediate post-tests at four strength levels by learning condition

		Massed				Spaced			
		productive recall	productive recall	productive recognition	receptive recognition	productive recall	productive recall	productive recognition	receptive recognition
nouns	mean	1.35	5.69	13.43	14.41	2.82	8.76	14.22	14.76
	median	1.00	6.00	14.00	15.00	3.00	9.00	14.00	15.00
	SD	1.23	2.94	1.86	0.98	1.58	3.34	0.92	0.52
	min	0	0	8	11	0	3	12	13
	max	4	14	15	15	6	14	15	15
verbs	mean	0.27	2.20	11.08	13.14	0.63	4.63	13.14	14.02
	median	0.00	1.00	12.00	14.00	0.00	5.00	14.00	15.00
	SD	0.53	2.35	2.79	1.78	0.86	2.92	1.74	1.31
	min	0	0	4	9	0	0	8	10
	max	2	9	15	15	3	11	15	15

Table 4.11 Descriptive statistics for word class scores in the delayed post-tests at four strength levels by learning condition

It is apparent that the means for all nouns are higher than the corresponding means for verbs. The patterns seen earlier are also replicated, in that the immediate test means are all higher than corresponding delayed test means, and the scores fall across the strength levels in the predicted pattern: receptive recognition > productive recognition > receptive recall > productive recall. Furthermore, mean scores from the spaced condition always exceed their massed condition counterparts. It is worth noting however that some of the differences between means are quite small and the median scores, reflecting as they do only the order of the scores, not their precise numerical values, show fewer descriptive differences. Medians of the immediate test exhibit no difference between nouns and verbs at the productive and receptive recognition strength levels in the massed condition and at the receptive recognition strength level in the spaced condition. The delayed post-test median scores show no difference between nouns and verbs at the productive and receptive recognition levels in the spaced condition. It is possible that the differences between nouns and verbs do not show themselves clearly at the easier strength levels due to the small SDs and very high scores at those levels, which resulted in a ceiling effect. The opposite is also noticeable at the highest strength level (i.e., productive recognition) where the scores are near or at zero (i.e., floor effect) in both learning conditions in the immediate and delayed post-tests.

A GZLM analysis was again conducted to test the significances of the main and interactive effects of the four factors involved: part of speech, learning condition, time, and strength (see Table 4.12). The analysis indicated many significant effects, which have been seen already, but the focus here is just on those effects that involve word class as main or interactive effect. Some effects which repeat findings that were seen in the earlier analyses are not revisited, such as a significant difference overall between immediate and delayed tests, between strength levels, and between spaced and massed learning. The most prominent result involving word class is the main effect of word class on scores, which is highly significant, although the size of the effect (judged from the Wald statistic) is lower than that of the main effects of time, strength level, or learning condition.

Still, this effect reflects a significant overall difference between scores for nouns and verbs, with the former higher than the latter regardless of time or learning condition or strength level.

With respect to interaction effects involving word class/part of speech, which is the main concern here, it was noticeable that none of the three results including interactions between word class and learning condition showed significant effects. That means that any effects of word class on scores were parallel in different learning conditions, so unaffected by the practice mode.

However, there were significant interaction effects involving word class and time and/or strength, without learning condition. The time by word class effect may be seen in Figure 4.5. It was noticeable that not only did verbs score lower than nouns, but scores for verbs also dropped between immediate and delayed tests further than those for nouns (i.e., regardless of learning condition). This effect was similar in size to that of the word class main effect (i.e., Wald statistic 78). This finding, therefore, indicated that, where word class is concerned, nouns are not only easier to initially learn but also easier to retain than verbs.

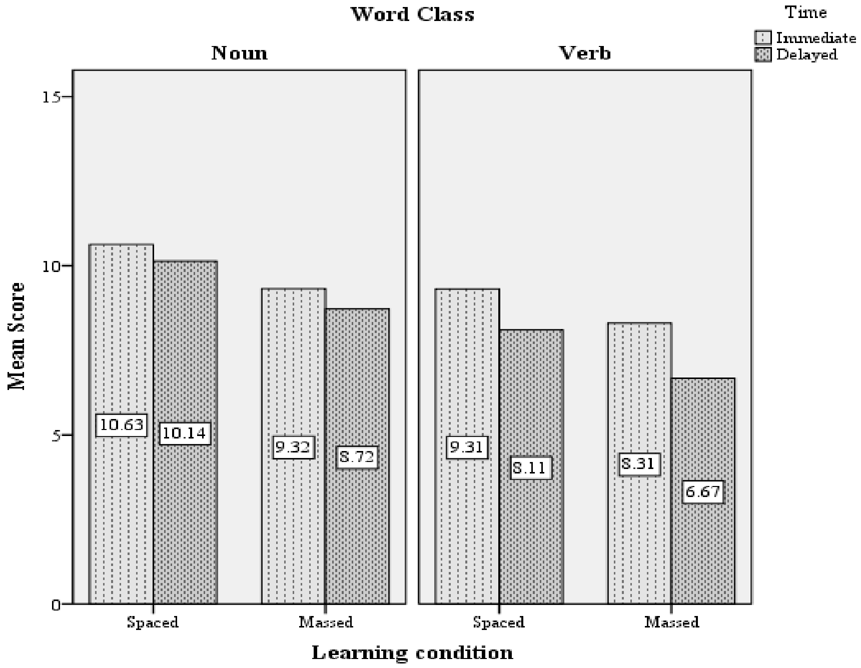


Figure 4.5 Mean vocabulary scores for word classes by Massed and Spaced conditions over Immediate and Delayed post-tests (combined strength levels)

Source	Wald Chi-Square	df	Sig.
Time	210.363	1	<.001
Word Class	77.970	1	<.001
Learning condition	83.980	1	<.001
Strength	482.814	3	<.001
Time * Word Class	77.218	1	<.001
Time * Learning condition	11.457	1	.001
Time * Strength	33.207	3	<.001
Word Class * Learning condition	.294	1	.588
Word Class * Strength	14.545	3	.002
Learning condition * Strength	27.394	3	<.001
Time * Word Class * Learning condition	.757	1	.384
Time * Word Class * Strength	13.003	3	.005
Word Class * Learning condition * Strength	3.163	3	.367
Time * Word Class * Learning condition * Strength	7.559	3	.056

Table 4.12 Overall GZLM analysis of Immediate and Delayed Post-test scores, by Learning condition, Strength and POS

The other notable interaction effect involving word class was that with strength level and time, regardless of learning condition. Although this interactive effect was also highly significant, the Wald chi squared was much lower than that of the preceding effect. This interactive effect shows that verbs not only fell more than nouns between immediate and delayed post-tests, but this fall further differed at different strength levels (see Figure 4.6). In particular, the difference between nouns and verbs in fall with time was greater for receptive recall (level 2) than the other strength levels. Furthermore, at level 1 (i.e., productive recall) the fall over time was unusually almost identical for nouns and verbs (i.e., around 1.1 words). A probable reason for this result was that receptive recall (i.e., level 2) was the level of knowledge where there was most space on the score scale for differences to show up. By contrast, the productive recall level scored quite low so suffered from floor effect while the productive recognition (i.e., level 3) and receptive recognition (i.e., level 4) scored high so experienced ceiling effect.

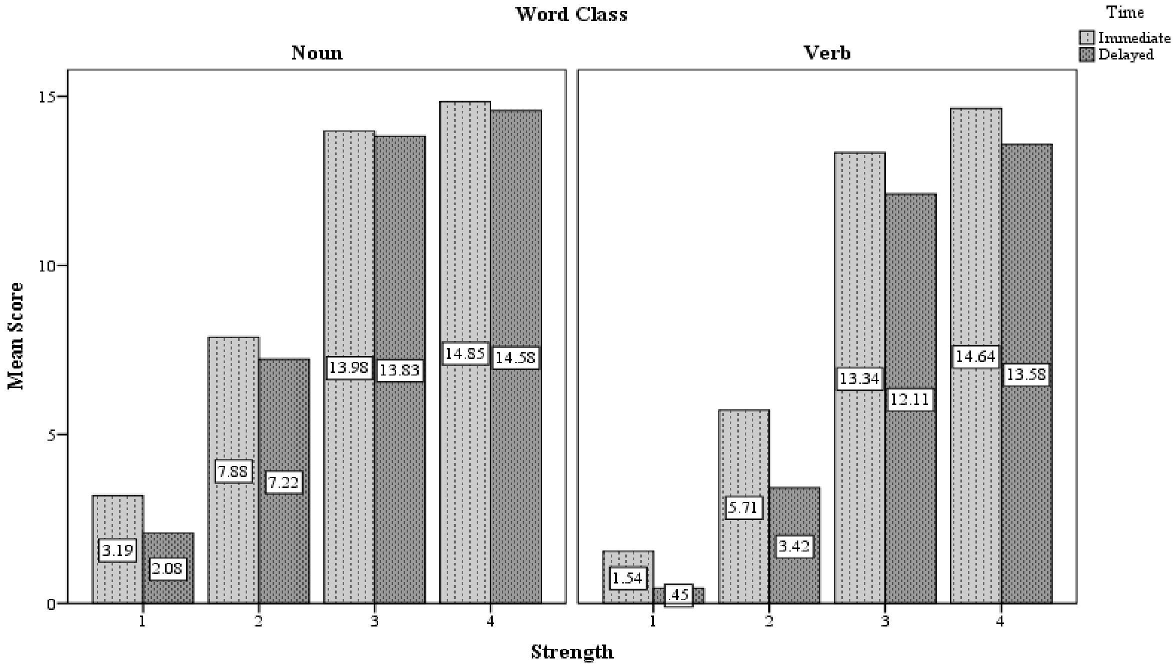


Figure 4.6 Mean vocabulary scores for Massed and Spaced conditions over Immediate and Delayed post-tests: each strength level separately.

4.6 RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed learning?

Only the responses to the closed questionnaire items are considered here since in response to the open-ended questions none of the participants provided any insights related to their perceived benefit of the two learning methods. A probable reason for the participants ignoring this part of the questionnaire could be fatigue due to the fact that the second questionnaire was administered immediately after the delayed post-test. The descriptive statistics of the questionnaire scores are reported in Table 4.13.

In order to answer RQ5a, the agreement ratings of students were considered in relation to the two sets of questionnaire items that had been deliberately included and which were confirmed as distinct by the factor analysis and Cronbach alpha coefficients (see Section 3.10.2). The two sets of questionnaire items respectively measure attitude/perception concerning two distinct dimensions of using the two types of learning (See Section 3.7.3). These two distinct dimensions of using the two types of learning were the perceived learning benefit itself, of various types, and aspects of reported engagement such as might enhance learning (e.g., attention and interest).

Looking first at the five items reflecting student perceived benefit of spaced over massed learning, it was found that many students remained neutral and did not indicate a preference for either practice type (rating 3) (see Table 4.14). Of those students who did express a preference (i.e., ratings 1 2 4 5), the majority preferred massed practice, except on item *d*.

Construct	Items	Mean	Median	Minimum	Maximum	SD
Learning	a. I memorize words better in spaced learning than massed learning.	2.98	3	1	5	0.95
	b. I memorize more words in spaced learning than massed learning.	2.76	3	2	5	0.69
	c. I retain words better in spaced learning than massed learning.	3.00	3	2	5	0.84
	d. I recall words better in spaced learning than massed learning.	3.20	3	2	5	0.82
	e. I learn words more quickly in spaced learning than massed learning.	2.90	3	2	5	0.62
Engagement	f. I can focus better in spaced learning than massed learning.	2.31	2	1	4	0.77
	g. I feel more motivated in spaced learning than massed learning.	3.20	3	2	4	0.68
	h. I feel less bored in spaced learning than massed learning	3.63	3	2	5	0.83

Table 4.13 Descriptive statistics for the questionnaire concerning learners' perceived difference between massed and spaced practice

Using the binomial test to see if those who responded positively significantly outnumbered those who responded negatively (see Table 4.14), a significant preference was found only on item 1b, where participants agreed that they thought massed learning led to better memorization. This picture was also reflected in the graph of the mean ratings (see Figure 4.7).

With respect to the benefit of each method for engagement, again many students chose the neutral middle agreement rating. Among those who expressed an opinion, however, there was a clearer division of opinion on two of the three items. A highly significant majority thought massed learning was better in terms of being able to focus and being less bored (see Figure 4.8). Overall, then, the students' opinion about what benefited them conflicted with the actual vocabulary score benefits they obtained.

Scale	**Items**	**Better spaced**	**Better massed**	**Neutral**	**p**
		n	n	n	
learning	a	14	17	18	.720
	b	5	18	26	.011
	c	13	15	21	.851
	d	14	8	27	.286
	e	5	11	33	.210
engagement	f	3	31	15	<.001
	g	17	7	25	.064
	h	2	24	23	<.001

Table 4.14 Tests of significant preference in responses to questionnaire items (Binomial test, omitting neutral responses)

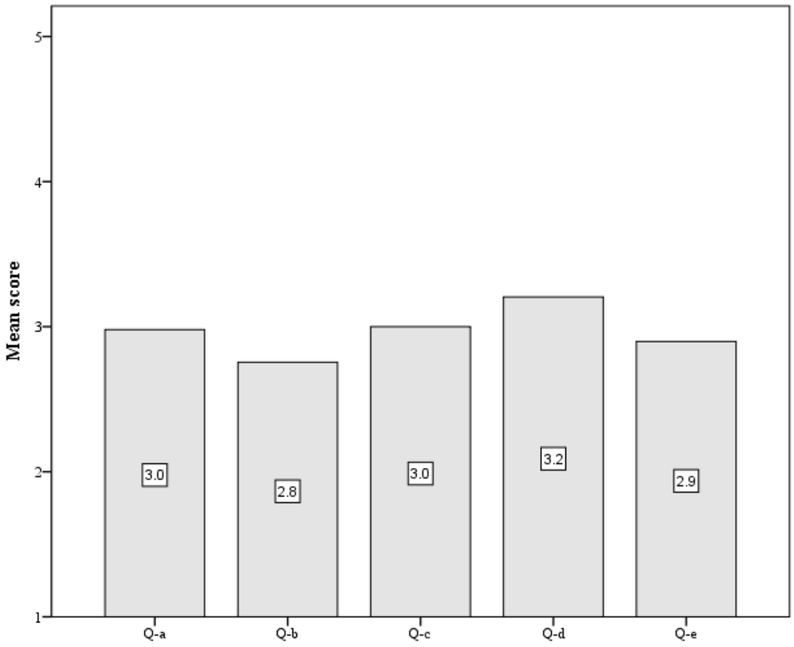


Figure 4.7 Mean ratings of students for perception of learning benefit (spaced high, massed low)

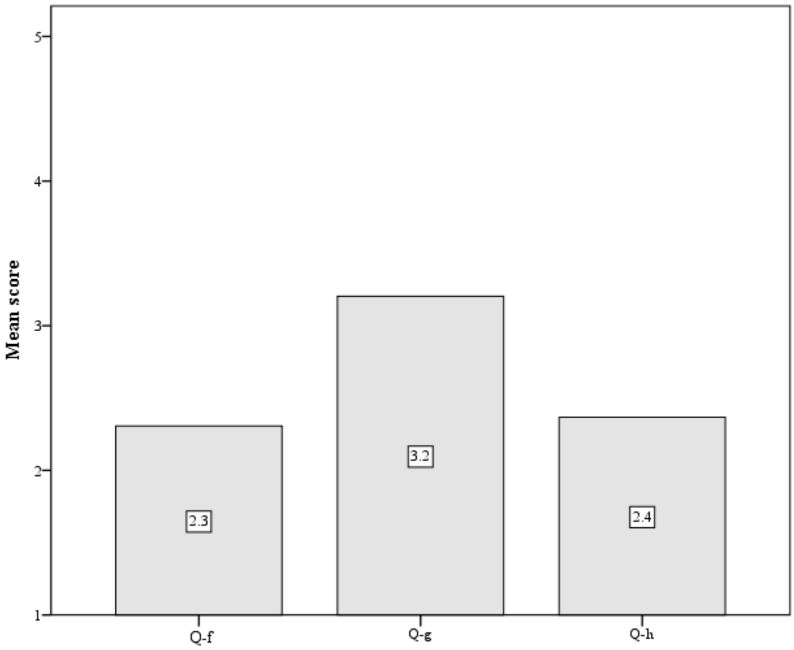


Figure 4.8 Mean ratings of students for perception of engagement benefit (spaced high, massed low)

4.7 RQ 5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test?

Spearman correlations between the participant rating scores established from the questionnaire and participant vocabulary scores at all levels of strength on both the immediate and delayed post-tests were calculated for each learning condition separately. Furthermore, correlations between questionnaire ratings and scores for degree of change of vocabulary knowledge between immediate and delayed tests were examined. In only one instance was any correlation significant. That was the relationship between degree of overall positive attitude to the learning benefit of spaced learning and the amount of immediate-delayed drop in productive recognition scores after learning in the massed condition (*rho* = -.396, *p* = .005).

The one isolated significant result made sense, in that one might expect that greater belief in spaced learning would be associated with less actual success with massed learning. However, the correlations in general do not indicate any systematic connection between a learner's attitude to or belief about the learning mode and their objective success with it.

4.8 Further exploration of the data

4.8.1 The levels of strength of lexical knowledge as an implicational scale

Although the four levels of lexical knowledge that the students practised and were tested on are not themselves a prime focus of this investigation, it was deemed interesting to further examine the data to test whether the four levels of vocabulary strength additionally constituted an implicational scale. In answering RQ1, RQ2, and RQ3 above, it has already been shown from the mean scores of the participants, and significance tests of the differences between them, that the data yields a clear order among the four vocabulary strength levels, and that the order is as predicted by Laufer and Goldstein (2004). The strength levels fall in an order of difficulty from the easiest to the hardest, regardless of practice condition or time when measured: receptive recognition > productive recognition > receptive recall > productive recall. Results in RQ1, RQ2, and RQ3, however, do not indicate directly whether the levels additionally form an implicational scale.

In order to confirm the implicational order of difficulty, not just the mean scores, but also scores for each individual participant on the tests for each item in each test condition must fall successively across the strength levels. Since at that level of detail (i.e., individual student scores for individual words in individual tests) the scores are simply 1 or zero (=correct or incorrect), one is essentially asking a series of questions of the following type: Is it always true that if students in a test get a particular word right at the productive recall level, they also get it right at all the other three supposedly easier levels? If the students get it right at the receptive recall level but not productive recall, do they also get it right at the two recognition levels which are supposed to be easier? If they get it right at the productive recognition level, do they also get it right at the receptive recognition level but not necessarily at productive recall?

The general principle of an implicational scale is that if the participants get an item correct at any point on the scale, they will also get it right at points lower than that in the implicational hierarchy (=easier strength levels, in the case of this study). Violations of that principle are then referred to as misfits (Hatch and Lazaraton, 1991). Table 4.15 illustrates this from the data of one of the students on one set of tests as recorded in the Excel data log. Table 4.16 shows the same data with the words on the side re-organised so as to demonstrate the implicational scale most clearly.

Table 4.16 shows the patterns of correct and incorrect answers across the four columns are consistent with the supposed order of difficulty in all cases except the words *prevail* and *isolate*. In those two instances the student gets the word right at the receptive recall level but not at the supposedly easier productive recognition level, so the responses violate the supposed order of strengths. In all other instances, however, harder levels are indeed got wrong and easier levels right, unless the student gets a word right at all four levels or wrong at all four levels. Thus, this example shows strong support at the micro level for the general finding from the means about the order of difficulty, and so supports the ordering as constituting an implicational scale. In other words, if this student gets a word right at a certain strength level it is possible to confidently (although not quite 100%) predict that the student will get it right at all easier strength levels too.

A statistic called the *coefficient of reproducibility* can be calculated from any such table, with a maximum possible value of 1 where there are no misfits. It is simply the total of non-misfitting numbers divided by the number of cells in the table (Hatch and Lazaraton, 1991). In Table 4.16 the number of misfits is 4, since for two words there is a 1 in the wrong place and a 0 in the wrong place. The number of cells is 30 x 4 = 120. Thus, the coefficient of reproducibility is (120-4)/120 = .967, which is of course very high. A further statistic called the *coefficient of scalability* is also often calculated, which adjusts that figure for the fact that the minimum possible value for a coefficient of reproducibility is not in fact 0 but varies depending on the pattern of results in a table. Thus, the coefficient of scalability rescales the coefficient of reproducibility onto a scale which does run from 0 to 1, so is easier to interpret. In the present example the coefficient of scalability is .811, which is still high.

Target Word	Productive recall	Receptive Recall	Productive Recognition	Receptive Recognition
debris	0	0	0	1
permit	0	0	1	1
trauma	0	0	1	1
escape	0	1	1	1
vendor	0	0	1	1
mentor	0	0	0	1
refuge	0	0	1	1
intent	0	1	1	1
sponsor	0	1	1	1
nominee	1	1	1	1
dignity	0	1	1	1
dilemma	0	0	1	1
pioneer	0	1	1	1
density	1	1	1	1
harmony	0	1	1	1
compel	0	0	1	1
battle	0	0	1	1
invade	0	0	1	1
comply	0	0	0	1
mutter	0	0	1	1
forbid	0	0	0	1
reward	0	0	0	1
regret	0	1	1	1
descend	0	0	0	1
concede	0	0	1	1
prevail	0	1	0	1
harvest	0	0	1	1
inherit	0	0	1	1
isolate	0	1	0	1
utilize	0	0	1	1
Total	2	11	22	30

Table 4.15 Student 7 in class 1 (immediate post-test, massed learning, word list B)

Target Word	Productive recall	Receptive Recall	Productive Recognition	Receptive Recognition
nominee	1	1	1	1
density	1	1	1	1
escape	0	1	1	1
intent	0	1	1	1
sponsor	0	1	1	1
dignity	0	1	1	1
pioneer	0	1	1	1
harmony	0	1	1	1
regret	0	1	1	1
permit	0	0	1	1
trauma	0	0	1	1
vendor	0	0	1	1
refuge	0	0	1	1
dilemma	0	0	1	1
compel	0	0	1	1
battle	0	0	1	1
invade	0	0	1	1
mutter	0	0	1	1
concede	0	0	1	1
prevail	0	1	0	1
harvest	0	0	1	1
inherit	0	0	1	1
isolate	0	1	0	1
utilize	0	0	1	1
debris	0	0	0	1
mentor	0	0	0	1
comply	0	0	0	1
forbid	0	0	0	1
reward	0	0	0	1
descend	0	0	0	1
Total	2	11	22	30

Table 4.16 Student 7 in class 1 (immediate post-test, massed learning, word list B), rearranged

It should be noted that the detailed misfits are not recoverable from the means. In the example above, the total scores of the student at each level fall in the predicted order, and that would be reflected in the mean scores for groups students, but such scores do not reflect the numbers of violations of the implicationality of the scale. Table 4.17 provides a possible but imagined example in which the total scores on each test at each level are the same as those above, but there are in fact nine violations of the implicational scale (i.e., 18 misfits), not just two, so this data would not support the presence of an implicational scale nearly so strongly as the data in Tables 4.15 and 4.16.

Target Word	Productive recall	Receptive Recall	Productive Recognition	Receptive Recognition
debris	0	**1**	*0*	1
permit	**1**	*0*	1	1
trauma	0	0	1	1
escape	0	**1**	*0*	1
vendor	0	0	1	1
mentor	0	1	1	1
refuge	**1**	*0*	1	1
intent	0	0	1	1
sponsor	0	1	1	1
nominee	0	0	1	1
dignity	0	1	1	1
dilemma	0	0	1	1
pioneer	0	0	1	1
density	0	1	1	1
harmony	0	**1**	*0*	1
compel	0	0	1	1
battle	0	**1**	*0*	1
invade	0	0	1	1
comply	0	0	1	1
mutter	0	0	1	1
forbid	0	0	1	1
reward	0	0	0	1
regret	0	**1**	*0*	1
descend	0	0	1	1
concede	0	0	1	1
prevail	0	**1**	*0*	1
harvest	0	0	1	1
inherit	0	0	1	1
isolate	0	**1**	*0*	1
utilize	0	0	1	1
Total	2	11	22	30

Table 4.17 Imaginary data with added misfitting correct responses bold and misfitting incorrect responses in italics

Tables such as 4.17 above contain 30 opportunities for an implicational scale of strength levels to be supported or not, one for each word. An inspection of all such tables from the data, which contain a total of 5880 opportunities (49x30x4), remarkably revealed that there are only 23 violations, which corresponds to 46 misfits. In other words, 99.8% of the responses of the students across the four strength levels were individually consistent with the predicted order of strengths. The coefficient of scalability could not be meaningfully calculated because of the many tables where there were no misfits at all, which raises technical problems due to division by zero.

Accordingly, there is exceedingly strong support for the strength levels constituting an implicational scale. Furthermore, there is no clear evidence for any effect on misfits of the other variables in the study: part of speech, time, or practice condition. Of the 23 violations of the implicational scale, 13 involved verbs and 10 nouns; also, there were 14 in the immediate post-test and 9 in the delayed post-test. Neither of those differences were significant (binomial test *p*> .05). More marked, though not quite significant (*p*=.093), was the division between items learned through massed practice (7 violations) and ones learned through spaced practice (16 violations). Possibly this greater difference can be explained by the fact that students learned more words in the spaced condition at higher strength levels than they did in the massed condition. Hence there was more opportunity in tests of words learned in the spaced condition for ‘misfitting’ sequences of scores to occur. In short, if there are more 1s in a table, there is more scope for some sequences of 1s to be interrupted by a 0.

There is additionally a deeper reason why it is interesting to check how far strength forms an implicational scale. In SLA research, a hierarchy of difficulty is widely interpreted as indicating the order of acquisition of that aspect over time. Famously when implicational scales were first used in SLA research in the grammatical morpheme studies of the 1970s (e.g., Andersen, 1978), precisely this claim was made. Tests of students suggested for example that, when assessed at some lower intermediate stage of their learning, they got the third person -*s* in English wrong consistently more than they got the progressive *-ing* wrong. This was interpreted as showing that they learn the *-ing* before they learn the third person *-s* which separate research suggested to be the case. In the current study, it is possible to claim that the result shows not just that students almost universally, regardless of the word involved, find the four strength levels successively more difficult to get right in tests (i.e., receptive recognition, productive recognition, receptive recall, productive recall), but that this also signals a universal order of learning vocabulary information. That is to say, that with very little exception, learners learn a word first at the receptive level, when tested in recognition mode, then progress to the productive recognition level, and so on. As noted above, such a claim cannot be supported by the mean scores alone.

Therefore, the results here constitute an interesting addition to our knowledge about vocabulary learning, with respect to the types of lexical information covered by the four strength levels. It should be noted that this analysis agrees with Laufer and Goldstein's (2004) finding of an implicational scale across the same four degrees of strength in their CATSS test (see Section 2.10.5.1) although the current analysis was conducted in a slightly different way. First, in the current study, the analysis was conducted on the entire data, whereas in in Laufer and Goldstein (2004) the analysis was conducted on results for a small number of randomly selected words. Second, the analysis was conducted using tables with words down the side rather than the participants down the side, as in Laufer and Goldstein (2004). The way that the analysis was conducted in the present study to test the implicational scale could be more informative (Scholfield, personal communication, 2019). In effect, Laufer and Goldstein (2004) asked: For each word (of the subset chosen) considered separately, do the responses support the proposition that each of the students learn the target words in ascending order of the supposed difficulty levels? By contrast, the analysis that was followed here answered the question: For each student considered separately, do the responses support the proposition that each of the words is learned by that student in ascending order of supposed difficulty levels? Since acquisition is something that actually happens within a student, for many words, rather than within a word, for many students, the method used in this study seemed more informative.

4.8.2 Item analysis

A further interesting aspect of the data to investigate was the scores for individual words. This exploration, however, is limited to the strength level of receptive recall because in the analyses that have been reported earlier in this chapter, it is apparent that the greatest level of score variation was able to emerge for this level. In essence, the participants did not tend to score very close to the ends of the score scale, as they often did at the other three strength levels. Hence, it is possible to claim that the most well differentiated and accurate information on individual word difficulty would be obtained by focusing on the receptive recall level.

	Time 1				Time 2			
	Spaced		Massed		Spaced		Massed	
Relatively easy	thread	84.6	thread	87.0	thread	88.5	thread	69.6
	breeze	84.3	breeze	82.6	breeze	84.6	sleeve	68.2
	palace	69.9	sleeve	78.3	palace	69.3	breeze	65.5
	sleeve	69.0	palace	73.9	sleeve	68.6	palace	65.2
	helmet	66.9	helmet	69.6	helmet	68.2	helmet	63.2
	defeat	65.4	hostage	56.5	defeat	65.4	defeat	47.8
	collar	65.4	insert	56.3	collar	65.1	collar	46.9
			condemn	54.5				
Relatively hard	verdict	38.5					donate	13.0
	thrive	37.6	verdict	26.1			foster	13.0
	donate	37.5	plunge	25.7	condemn	30.8	exhaust	13.0
	persist	37.0	stumble	25.1	persist	29.9	texture	8.7
	disturb	36.8	vaccine	21.7	dictate	26.9	persist	8.5
	insert	34.6	texture	21.4	exhaust	23.1	exploit	6.4
	exploit	34.3	persist	20.0	foster	19.2	dictate	5.9
	foster	26.9	exploit	19.7	exploit	15.4	provoke	4.3

Table 4.18 List A top and bottom scoring words at the receptive recall level (% correct)

Accordingly, the six top and bottom scoring items in the receptive recall tests for each time, learning condition and word list were identified (Tables 4.18 and 4.19). Where further items had scores that tied with that of the sixth item listed, they are displayed as well. A number of interesting observations can be made on the basis of this data.

	Time 1				Time 2			
	Spaced		Massed		Spaced		Massed	
Relatively easy	debris	95.7	permit	46.2	permit	82.6	refuge	34.6
	permit	87.0	debris	42.3	intent	78.3	debris	34.5
	intent	78.3	trauma	42.3	debris	73.9	escape	30.8
	escape	73.0	refuge	42.1	escape	73.4	sponsor	30.6
	regret	73.9	escape	38.5	mutter	65.2	permit	26.9
	trauma	69.6	vendor	38.2	regret	65.0	intent	26.8
	mentor	69.0	sponsor	38.0			harmony	26.4
			nominee	37.6			trauma	26.1
			dignity	37.5				
			pioneer	37.5				
			density	36.7				
			harmony	36.6				
Relatively hard					compel	34.8	vendor	7.7
	compel	39.1			invade	34.6	compel	7.5
	invade	39.0	mutter	23.5	harvest	33.9	dignity	7.4
	prevail	39.0	concede	23.1	inherit	33.8	density	7.1
	utilize	38.9	prevail	23.0	isolate	33.8	comply	7.1
	vendor	34.8	inherit	23.0	utilize	33.0	isolate	6.8
	comply	34.5	forbid	19.2	vendor	30.4	prevail	3.8
	harvest	34.5	descend	18.7	comply	30.1	concede	3.6
	concede	30.4	utilize	18.5	concede	30.0	utilize	3.5

Table 4.19 List B top and bottom scoring words at the receptive recall level (% correct)

In each word list there is a clear set of up to five words that prove consistently high or low scoring regardless of learning condition or test time. At the top/easier end in list A, there are the following words: *thread, breeze, palace, sleeve,* and *helmet* while in list B it was *debris, permit, escape*. At the bottom end in list A are *exploit* and *persist*, while in list B *concede* and *utilize*.

In order to explain these findings whether the position of a word in the exercises where it was practiced played any role was first considered. It should be noted that the order of the occurrences of all the target items in the four exercises was randomized using Research Randomizer v4 (2013). Nevertheless, it might be possible to argue that some items were easy for the participants because these items consistently chanced to appear closer to the beginning of each exercise than other items that came closer to the end of each exercise and proved harder. The position of all the top and bottom scoring items in the four exercises was therefore examined (see Table 2.20).

		Item	Exercise				Mean	SD
			1	2	3	4		
Top scoring items	List A	thread	2	14	1	24	11	11.3
		breeze	18	7	18	11	14	5.4
	List B	debris	4	27	3	25	15	13.0
		permit	14	9	28	10	16	8.6
Bottom scoring items	List A	exploit	16	19	16	22	18	2.9
		persist	13	12	13	21	15	4.2
	List B	utilize	28	26	17	11	20	8.3
		concede	29	23	7	20	20	9.3

Table 4.20 Descriptive statistics for the order of occurrence for each top and bottom scoring item in four exercises.

Next, the positions where a word could appear was divided into three areas: near the start of the exercise (position 1-10), in the middle (position 11-20), or near the end (position 21-30). As can be seen in Table 4.20, the top scoring items occurred 44% of the time closer to the beginning, 31% of the time during the middle, and 25% closer to the end of the exercise sessions. Only one item (i.e., *concede*) from the bottom scoring items however occurred closer to the beginning of exercise three, whereas the remaining instances of practicing the words occurred 50% during the middle or 44% near the end of the exercise sessions. Thus, it is possible to claim that the participants found it easier to learn *thread, breeze, debris,* and *permit* because these items appeared 63% of the time during the first half of each exercise, as opposed to *exploit, persist, utilize,* and *concede*, which appeared almost 69% during the last half of the exercise sessions.

Part of speech provides a second contribution to explaining differences between words. As might be expected from the general result in Sections 4.6, the highest scoring items listed in Tables 4.18 and 4.19 are mostly nouns (or were in this study treated only as nouns, e.g., *defeat*), and the low scoring items are verbs. There are however some exceptions. The verbs *insert, condemn,* and *regret* sometimes feature among the easiest items, while the nouns *texture, verdict, vaccine, vendor, density,* and *dignity* appear among the most difficult items.

Thirdly, it is also noticeable that list A exhibits very high consistency across conditions in the words that proved easiest to learn (at the receptive recall level). The average number of occurrences in the sub lists of items across the four conditions in Table 4.18 is 2.9 for the more learnable words but only 1.8 for the less learnable words. For instance, there are four items among the easiest items that appear in all four exercises (*thread, breeze, sleeve, helmet*) but only one such word among the hardest items (*persist*). On the one hand, this difference between the more and less learnable words suggests that the time of testing (i.e., immediate, and delayed post-tests) and the practice conditions (i.e., massed, and spaced practice) make no great difference to which words the students find relatively easier to learn. On the other hand, however, time and practice conditions do seem to create more diversity in the words that prove hardest to the students. It is not entirely clear why these factors affect words differently and curiously this effect was not found clearly in List B, where the average rate of occurrence of words in different conditions was 2.1 for the top scoring words and 2.2 for the lowest scoring ones. This is something that deserves further research.

A further observation is that, comparing the two times when the tests were done, it is also notable that not every individual word falls in score between times. This applies most strikingly to the top scoring words in the spaced condition in list A (Table 4.18) which all occur in the top scoring list both in the immediate and delayed post-tests. Of these words*, breeze, palace, sleeve, helmet, defeat* and *collar* all maintain identical mean scores, while *thread* actually increases its score in the delayed test. Once again, this effect is not so strong for list B where *intent* and *escape* maintain their relatively high scores, but the remainder drop between immediate and delayed test.

By contrast, with the one exception of *concede*, all spaced scores for the lowest scoring words fall between immediate and delayed test. In addition, all the massed scores for words fall between times, regardless of whether they are high or low scoring. Still, this fall of scores between the immediate and delayed post-test further supports what was noted in Section 4.5 above, that in the current study where more is initially learned, the drop between immediate and delayed post-tests is smaller than that where less is initially learned, especially in the spaced condition.

Finally, from the general results in earlier sections it is not surprising to find that in most cases the scores for words learned in the massed condition are lower than those for the same words learned in the spaced condition. However, it is instructive to see that this is not always the case, and this occurs in List A, which once again stands out from list B, and again for the highest scoring words rather than the lowest. In Table 4.18, it can be seen that in the immediate post-test *thread*, *palace, sleeve,* and *helmet* actually score higher after the massed practice than the spaced. By the time of the delayed post-test, however, that situation has reversed due to the far better retention consequent upon the spaced learning.

Chapter Five - Discussion

The following sections provide a discussion of the results in this study which were presented in chapter 4, in the order of each research question.

5.1 RQ 1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?

The immediate test scores suggested that spaced practice led to higher initial vocabulary learning than massed practice at only three strength levels out of four, namely productive recall, receptive recall, and productive recognition. At the lowest strength level (i.e., receptive recognition) the differences in vocabulary gains between massed and spaced conditions were hardly noticeable, although it should be noted that spaced learning produced higher gains descriptively. As mentioned earlier (see Chapter 4, Section 4.2), the lack of difference at the receptive recognition level was explainable by the fact that scores for receptive recognition were often 100% correct or nearly so, so ceiling effect was at work limiting the scope for any difference between learning conditions to show itself.

These results do not support findings in earlier studies which described both massed and spaced practice as equal in effect on initial vocabulary gains (e.g., Bloom & Shuell, 1981; Nakata & Webb, 2016). There are three possible reasons that could explain the differences between the findings in the current study and previous research.

First, the type of measurement that was used in the previous studies did not measure the effect of massed practice and spaced practice on vocabulary knowledge at four strength levels as in this study. For example, Bloom and Shuell (1981) assessed vocabulary knowledge at only the productive recall level by using an L1-L2 translation test, while Nakata and Webb (2016) assessed vocabulary knowledge at the receptive and productive recall levels through the use of an L2-L1 translation test and an L1-L2 translation test, respectively. Hence those studies provide no comparable results to the current study for the receptive recognition level. In the present study, however, the strength levels that are comparable with these studies exhibited quite different results. For instance, receptive recall was the level of knowledge that actually exhibited the greatest difference between learning conditions, of more than four words learned in the immediate post-test. Hence it remains surprising that Bloom and Shuell (1981) and Nakata and Webb (2015) found no difference.

Another possible explanation however could be that, unlike in those studies, the participants in the present study had to learn each target word at four levels of vocabulary knowledge. The participants were not just repeating exercises at the same strength level. It is possible that practicing words at different strength levels, even if they were not fully acquired at those levels, further enforced learning these words at other strength levels. Accordingly, it could be possible that if other strength levels had been targeted during the treatment stage and then assessed in those other studies, they could have obtained more similar results to the ones in this study and shown an advantage in vocabulary gains by using one practice type over the other. Bloom and Shuell (1981) for example provided practice exercises only at the productive recognition and productive recall levels, before a test of productive recall. If they had also included practice exercises at the receptive levels first, possibly the impact of the learning conditions would have showed up more strongly in the initial post-test scores.

Third, the participants' prior knowledge of the target words might have played a role in the contradicting findings. For example, Bloom and Shuell (1981) did not examine the participants' prior knowledge of the target words. It is generally advocated in research into vocabulary acquisition to establish “what vocabulary knowledge exists at a point in time (usually before an experimental treatment), and then establishing what the state of knowledge is at a later point” (Schmitt, 2010, p. 179). If it is taken into consideration that the participants in Bloom and Shuell's (1981) study were second-level French learners, it is possible that some of the target words were already known by the participants which could have influenced their findings. If there is no pre-test, and part of what was tested in an immediate post-test was in fact already known before the study, then it is quite likely that the effect of different learning methods employed just before the immediate post-test will not appear clearly in those test results. On the other hand, Nakata and Webb (2016) did in fact examine the participants’ prior knowledge of the target words. However, the type of measurement they used examined vocabulary knowledge at receptive recall level (i.e., L2-L1 translation test). Therefore, it is possible that the participants in their study may have known the target words at lower strength levels (e.g., productive recognition, receptive recognition), which could have provided an advantage to certain words that were known at a lower strength level over other words that were not known at all.

Fourth, the participants in Bloom and Shuell (1981) were English-speaking learners of French, while in the present study the participants were Arabic-speaking learners of English. It is quite possible that the degree of similarity between English and French may have played a role in how massed practice and spaced practice affected vocabulary learning. Again, this is a factor potentially affecting scores but independent of the massed-spaced distinction. In particular, the participants in Bloom and Shuell (1981) were learning French occupation names which very commonly have similarities with English occupation names (e.g., French/English: *chauffeur*/*chauffeur*, *électricien*/*electrician*, *pilote*/*pilot*, *journaliste*/*journalist*). Note, however, it was not possible to verify this assumption since Bloom and Shuell (1981) did not publish the wordlist that was used in the study.

While the spaced learning effect evidence in the current study is not found in the above studies close to it, it should be recognised that in the massed-spaced literature in general (across disciplines) it is common to find an immediate post-test effect of spaced instruction/learning. In a meta-analytic survey of many such studies, Donovan and Radosevich (1999) found an overall effect size of .46 for what they term 'acquisition' (i.e., knowledge tested immediately after the last practice session). They measured effect size with Hedges' g coefficient, and the equivalent figure for the present study, at the knowledge level which showed the greatest massed-spaced difference (i.e., receptive recall), is .117. This shows that, viewed in the much wider context of massed-spaced research in general, the immediate differential impact of the conditions, although significant, is quite modest.

A further finding of the present study is that the strength scores exhibited a rise in scores across the four different levels of strength in the way predicted by Laufer & Goldstein (2004) but not by Laufer et al. (2004). It is possible that this finding is congruent with Laufer & Goldstein (2004) because the target participants in this study and in Laufer & Goldstein (2004) were EFL learners, while the participants in Laufer et al. (2004) were ESL learners. Therefore, this finding confirmed the order of the degrees of strength of vocabulary knowledge among EFL learners, which, regardless of the spaced-massed distinction, further supports the validity of the VATS as a measuring instrument in this study.

Finally, to my knowledge, this study was the first to examine the difference between massed and spaced practice on EFL vocabulary learning at four levels of vocabulary knowledge. The differences between learning conditions were greater than three words at the two higher strength levels (i.e., productive recall and receptive recall), while the two low strength levels (i.e., productive recognition and receptive recognition) had a difference of less than one and a half words. In other words, the effect of the learning method in terms of the difference in the number of words learnt is at least twice as much at the more demanding recall strength levels than at the less demanding recognition strength levels. A probable reason for the differences being small for the productive recognition and receptive recognition levels is that the task of learning the words at these levels of knowledge proved not particularly challenging regardless of the learning condition. Hence, the data exhibited ceiling effect in that scores in the massed and spaced learning conditions were very near 100% at the receptive recognition and productive recognition levels, and for that reason there was no space for large differences between the two conditions to show themselves.

It is worth pointing out that, in real life vocabulary use in speech and writing, people are using language in a way more like the recall than the recognition levels as tested in this study. Normally one has to retrieve the word for one's intended meaning when speaking or writing, without any multiple-choice support, similar to being offered an L1 definition and simply having to supply the L2 equivalent, as in what is termed *recall* in this study. Again, when one sees or hears a word (reading, listening) one has to retrieve the meaning without multiple choice support, although admittedly other forms of support may be present in the context (e.g., semantically, or pragmatically related words). In real life, there are only rare instances where one is offered a multiple-choice range of word forms or meanings, containing the correct one, at the point of word use or understanding. That tends only to be the case in classroom exercises and tests. Hence, it may be concluded that it is particularly helpful for learners if spaced learning differentially assists the more challenging task of recall and of less interest if it helps recognition, and indeed that is what seems to occur.

It should be noted that the effect size of the spaced condition was not huge. Its best performance was in the receptive recall condition where it led to a little over four more words being initially learned than in the massed condition, with a modest effect size g=.117. Still, however, this is pedagogically an advantage worth gaining.

5.2 RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test?

The present study supported and extended the findings of Sobel et al. (2011) and Goossens et al. (2012) who found that spaced practice led to higher vocabulary retention over a period of five weeks than massed practice. It should be noted however that, in their studies, the participants were primary school students who were native speakers of the language of the words being learned (respectively English and Dutch), while in the present study the participants were EFL university students. Therefore, the results in the present study have remarkably demonstrated that spaced practice could be equally effective in learning vocabulary at extremes of the educational continuum.

While the present study supports the common finding in other studies that spaced learning is more effective than massed learning, it does not do so as strikingly as some of those studies. Sobel et al. (2011) for instance, in a delayed receptive recall test, obtained scores for the spaced condition that were three times those for the massed condition with a difference 177% greater than the massed score. This is not exactly replicated by the current study where at that strength the spaced condition yielded mean scores a little over twice the size of those in the massed condition with a difference only 101% of the mean massed score. However, as indicated earlier, Sobel et al. (2011) sampled quite a different population from this study (i.e., primary level native speakers). Furthermore, a number of methodological criticisms of Sobel's study was raised (see Section 2.9.4) relating to the small number of words, lack of checking of prior knowledge, and so forth. Notably, it has been remarked in a meta-analytic study of 63 massed-spaced studies that "higher effect sizes were found in studies with low methodological rigor as compared with those studies higher in rigor" (Donovan and Radosevich, 1999, p. 795).

Looking again at the meta-analytic findings of Donovan and Radosevich (1999) across massed-spaced studies in multiple disciplines, it can be seen that they record a mean effect size for retention of g=.52. This compares with .240 for the best difference in the data of the present study (at receptive recall level). Thus, the data of the current study matches the meta-analysis in that it shows a stronger effect of spaced instruction on retention than on initial acquisition (cf. Section 5.1). However, on a broad view, the size of the effect in this study, although significant, is not massive.

Once again, the current study has provided a new finding in relation to the effect of massed and spaced practice since, unlike those two cited studies, the current study tested the retention of EFL vocabulary at four levels of vocabulary knowledge strength. After a four-weeks retention interval, compared with the initial baseline of zero knowledge, scores increased again successively in each learning condition across the four strength levels from the most demanding to the least demanding (i.e., productive recall < receptive recall < productive recognition < receptive recognition). Once again, the two recognition levels stood out as the easiest, with scores quite close to each other for the receptive and productive options, while the more demanding recall levels scored more distantly lower and less close to each other.

The differences in scores between learning conditions at each of the strength levels were generally greater than in the immediate post-test, but no longer larger at the two higher strength levels than at the two low strength levels (as in the immediate post-test). In fact, the most demanding level (i.e., productive recall) exhibited a difference of only 3.3 words between the learning conditions, which was the same difference as was seen in the immediate post-test and similar to the difference for delayed productive recognition at 3.5 words. In other words, while generally the delayed post-test results suggested bigger differences produced by the two learning conditions than in the immediate test results, this difference was no longer more the case at both of the more demanding strength levels but primarily just for receptive recall, which stood out from the rest of the levels with a learning condition difference of over 7 words. As suggested previously (See Chapter 4, Section 4.3), this could be due to the receptive recall level being the one that generated scores that were most distant from the floor and ceiling of the score scale. Hence, it had more 'space' on the score scale in which variation could emerge. Nevertheless, the higher differences found at all three lower strength levels, in the delayed post-test, compared to the immediate post-test, demonstrated that spaced practice was especially effective in enhancing longer term retention.

The greater enhancement of delayed than of initial learning by the spaced condition has also been found in other areas of study, such as L2 reading comprehension (e.g., Rawson & Kintsch, 2005), L2 grammar (e.g., Bird, 2011; Miles, 2014, Nakata & Suzuki, 2018), computer-assisted language learning (e.g., Lindsey, Shroyer, Pashler, & Mozer, 2014), and mathematics learning (e.g., Rohrer, Dedrick, & Stershic, 2015). Hence, the current study may be seen as confirming in a less well researched area a universal characteristic of the massed-spaced distinction.

5.3 RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength levels between the immediate post-test and a 4-week delayed post-test?

Although the immediate and delayed retention scores were examined separately, such measures can “only give a snapshot of vocabulary knowledge and cannot inform about the dynamic and incremental nature of the learning process” (Schmitt, 2011, p. 155). In order to access that dynamic nature, the current study therefore also examined the changes between the immediate and delayed post-test.

To my knowledge, this study in fact is the first classroom-based spacing effect study to measure the effect of massed and spaced practice on EFL vocabulary learning using both immediate and delayed post-tests after the end of the treatments. Consequently, it was possible to obtain more information on the effects of massed and spaced practice by comparing short-term and long-term classroom vocabulary learning.

In general, time, in the form of discontinued use, negatively affected the scores after both massed and spaced practice. This was expected due to the considerable evidence in the literature which suggested that forgetting would occur once repetition and exposure to the target items stopped (e.g., Horst & Meara, 1999; Milton, 2009; Alharthi, 2012). Most importantly, massed learning not only produced lower scores than spaced learning on the immediate post-tests, but also yielded a greater fall in scores than spaced learning did between immediate and delayed post-tests, except descriptively at the least demanding strength level. At the hardest strength level (i.e., productive recall), the differences in the loss of information became harder to distinguish, which was perhaps due to the greatest difficulty of the knowledge to be learnt at this level. In general, the scores at the productive recall level were quite low, so floor effect made differences less easy to appear. It was noticeable that descriptively the difference in vocabulary loss was greatest at receptive recall strength level, which was also the level where scores were more often obtained that were distant from the extremes of the scale (i.e., floor = zero, ceiling = 30).

Further exploration of the immediate-delayed changes in vocabulary scores threw detailed light not only on the interaction of the effect of time with learning condition, but also of time with the recognition-recall opposition in the four knowledge levels that were considered in this study. The fall in scores was in fact greatest for the recall tests. There was, however, no significant interaction effect of the receptive-productive opposition with time.

Since it was not possible to find other massed-spaced classroom vocabulary studies that measured learning on two occasions (i.e., immediate post-test and delayed post-test) and considered the same variables in the current study, the findings here are new and cannot be compared with other research.

5.4 RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed?

As previously reviewed in Chapter 2 (See Section 2.3.2 and Section 2.6.1), there have been mixed findings for whether word class/ part of speech has a clear effect in general on the difficulty or ease of vocabulary learning, regardless of consideration of the precise learning method. There have been however some studies of word class in massed and spaced learning which recorded an advantage for nouns. For example, Childers and Tomasello (2002) examined the effect of massed and spaced practice on receptive and productive learning of nouns and verbs among two-year old native English-speaking children. The results suggested that spaced practice only benefited productive knowledge, where children learned nouns three times as much as verbs. By contrast, the results in the current study indicated that the effect of massed and spaced practice was the same on receptive and productive vocabulary knowledge for both nouns and verbs, although there was a marked overall advantage to nouns.

Earlier studies (e.g., Sobel, et al., 2011; Goossens et al., 2012) did not examine the effect of massed practice and spaced practice on different parts of speech in L2 learners, so the findings here are new results which cannot be directly compared with any earlier ones for L2 learners, only L1 children. The current study found (like Childers & Tomasello, 2002) that verbs did score lower than nouns, but (contrary to that study) that this was true for both massed and spaced practice and both receptive and productive levels of knowledge. Furthermore, scores for verbs also dropped further than those for nouns between the immediate and the delayed post-tests, especially verbs that were tested at the receptive recall strength level.

Given that verbs seemed to be at a disadvantage in other ways compared with nouns (e.g., in both initial learning and retention), it was encouraging to find that learning verbs was not affected differentially by the type of practice used. This finding has pedagogical implications since teachers might be less encouraged to use spaced practice if it had emerged that this type of practice would only assist the learning of one part of speech rather than another.

In more detail, scores not only fell more for verbs than nouns between the immediate and the delayed post-tests, but this fall of scores further differed at different strength levels which, to my knowledge, nobody has considered before in relation to word class as a variable in vocabulary learning. There was an exceptionally high fall of scores over time for verbs at the receptive recall level and a lower than average fall for verbs over time at the productive recall level, which departed from the general pattern. A possible explanation for this finding, however, was that floor effect might have affected verbs at the most demanding productive recall level, reducing the possibility for scores to fall further over time than they did, while lack of proximity to floor or ceiling might have allowed the receptive recall scores for verbs to fall more than average. Floor or ceiling effect has been found in other spacing effect studies such as Childers and Tomasello (2002), Seabrook et al., (2005), Ambridge, Theakston, Lieven, and Tomasello (2006), Rohrer (2009), and Walsh et al., (2018). It is a feature of such studies which is hard to avoid, although it would be desirable if it did not occur, so as to allow the real difficulties of the words, the learning modes, and the tests/tasks to emerge without the limits of the score scale affecting the results.

The overall advantage of nouns over verbs in learning tasks has been explained by Maguire, Hirsh-Pasek, and Golinkoff (2006, p. 367) through the assumption that “mapping from action or mental state to word is considerably more challenging than mapping from object to word”. The deeper reason for that mapping difference however remains unclear. Some have suggested that the key factor is that words with concrete meanings are easier to learn than those with abstract ones because they are imageable (Gairns & Redman, 1986; Ellis & Beaton, 1993; de Groot, 2006). For example, a word such as *helmet,* which can be touched and seen, could be easier to learn than words such as *consent* or *verdict*. Accordingly, since nouns more often have concrete meanings than verbs do, it is arguably possible that for that reason nouns are easier to learn (Schwanenflugel, 1992). In the present study the result cannot however be simply attributed to nouns having concrete meanings and verbs abstract ones, since most of the nouns denoted concrete objects and most of the verbs denoted physical actions (e.g., *descend*, *mutter*, *stumble*, *shrink*). Therefore, most of the words chosen in this study were imageable. Thus, the current study confirms an advantage for nouns but does not support the explanation that this is due to their greater concreteness and imageability.

Another interesting observation of the word class results is that the word class that was associated with greater initial learning (i.e., nouns) also suffered the least amount of attrition. This runs counter to what some studies indicated (e.g., Keijzer, 2007; Alharthi, 2012), which is that the people or conditions associated with the greatest initial learning have, in a sense, more knowledge available to lose, and perhaps for that reason go on to lose more by the time a delayed post-test takes place than those who initially learned less, so had less to remember. The reason for the difference between nouns and verbs in this regard could however be attributable to the rather different nature of the research in those studies. The studies cited above were not studies like the current one, where learning occurred during a researcher intervention, and where retention was followed up over a month. Rather they were studies where the initial learning had occurred without special intervention, either during normal classroom instruction or natural immersion acquisition, and the retention was tracked over spans of years rather than weeks.

5.5 RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed learning?

With respect to the result for the learning benefits, it was discussed above (See Section 5.3) that the participants actually learnt significantly better in the spaced condition than in the massed condition. However, it also emerged (See Section 4.6) that in terms of the participants' attitudes, they were either indifferent as to which learning condition they preferred, or, if anything, regarded the massed approach as superior. This finding, therefore, suggested that these learners did not have well developed metacognitive ability in the area of their own language achievement. Basically, learners could not judge correctly where they learn more or less successfully. This issue has been found in other spacing effect studies (e.g., Dunlosky & Nelson, 1994; Simon & Bjork, 2001; Kornell & Bjork, 2008; Kornell, Castel, Eich, & Bjork, 2010; Logan et al., 2012).

It is not unusual in the language learner literature in general for there to be a mismatch between attitudes that students have to any educational treatment and their actual measurable success. For example, Garrett, Griffiths, James, and Scholfield (1994) found that while students liked the introduction of use of L1 rather than L2 to brainstorm before writing in L2, in fact this made little difference to the quality of what they wrote. In the realm of vocabulary, Lin (2014) found that while students' attitudes to using classroom group work rather than individual work to learn vocabulary fell over the period of the intervention study, the students’ success in learning vocabulary through group work actually improved.

With respect to the finding for engagement, it was perhaps understandable that the massed condition would be judged the one where it was easier to focus, precisely because all the information was delivered in one extended session rather than in fragments at different times. For the same reason it made sense perhaps that spaced learning might have appeared more boring to participants since it involved revisiting the same words on many occasions. Overall it was interesting therefore that the very feature that made spaced learning more effective, the fact that it involved repetition on separate occasions, may at the same time have made it less attractive to students due to the possible boredom and perceived breaking of concentration that this entailed.

5.6 RQ 5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test?

It emerged that there was almost no relationship between individual student perceptions of the benefits of the two approaches and how well they performed with them. This result, therefore, suggested that spacing the learning of vocabulary was equally beneficial to all students, irrespective of whether the students preferred the spaced practice or not over the massed practice. In the area of vocabulary learning through massed or spaced means, to my knowledge, this precise correlational issue has not been pursued before, hence this constitutes a new finding.

Finally, it is worth pointing out that the lack of a relationship between attitude and success is a good finding in that if it had emerged that spaced learning was only really successful with those students who had a positive attitude to it, I could not so easily argue, as I will in the Conclusion chapter, that spaced learning should be widely adopted. In general, classroom innovations are more desirable if they do not depend for their success on the preference of students, or on their wider characteristics such as individual learning styles.

Chapter Six - Conclusion

6.1 Introduction

This study had the aim of investigating whether massed practice or spaced practice can better enhance classroom vocabulary learning among Saudi EFL learners at Taif university. It has, however, much broader interest and importance since, to my knowledge, the current study is the first to examine the effect of massing and spacing on EFL vocabulary learning in authentic classroom settings at university level in general, not only at a Saudi university. In addition, this study is the first to investigate the spacing effect in EFL classrooms by establishing baseline information through testing prior target-word knowledge before teaching these words and finally testing the participants' short-term and longer-term vocabulary knowledge change at four different strength levels. Furthermore, only one earlier study (i.e., Childers & Tomasello, 2002) examined the effect of massed and spaced practice in relation to word class, and that concerned quite a different population from the present study, of two-year-old native speaker children.

The participants in the current study had to learn new vocabulary by completing four vocabulary learning exercises either in one classroom session (for the massed practice condition) or one different exercise at a time over four weeks of an equally distributed schedule (for the spaced practice condition). Furthermore, the effects of these two learning conditions were assessed in relation to four vocabulary knowledge strengths, word class, and learners' perceived benefits of vocabulary learning using massed practice and spaced practice.

This chapter begins by summarising the conclusions reached in Chapter 5 as they relate to each RQ. It then addresses some of the limitations of the present study. Afterwards, since the current study examined the effectiveness of the spacing effect inside an actual EFL classroom environment, the pedagogical and Second Language Acquisition theory implications of the study will be discussed. The final section will make recommendations and suggestions for future research on the spacing effect in EFL vocabulary learning.

6.2 Summary of findings

This section presents a summary of findings related to the different variables in the current study which may or may not have an impact on the retention or attrition of vocabulary knowledge among Saudi EFL First-year undergraduate students. These comprise the variables: type of instructional treatment (two types of practice; i.e., massed practice and spaced practice), aspects of lexical knowledge acquired (i.e., receptive recognition, productive recognition, receptive recall, productive recall), part of speech of word learned, retention interval (periods between pre-test, immediate post-test and delayed post-test), and two individual difference learner variables (i.e., learners' perceptions of the effectiveness of massed and spaced practice, and of the engagement associated with each).

6.2.1 The effect of repetition learning with or without spacing on immediate acquisition of vocabulary knowledge at four strength levels

RQ1: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on an immediate post-test?

H1: Massed and spaced practice have equivalent effects on vocabulary learning in the form of initial vocabulary knowledge gains (e.g., Bloom & Shuell, 1981).

Earlier attempts to examine the difference in effect of massed and spaced practice on EFL classroom vocabulary learning did not assess initial gains of vocabulary knowledge immediately after the completion of the massed or spaced practice (e.g., Sobel et al., 2011; Goossens et al, 2012). However, some cognitive psychology research suggested that massed and spaced practice have equivalent effects on initial vocabulary knowledge gains (e.g., Bloom & Shuell, 1981).

Therefore, anything that this research does to shed light on the difference in effect of the two practice methods on initial gains of vocabulary knowledge has implications for EFL vocabulary teaching (see Section 6.5).

The results of the immediate post-test revealed that spaced practice yielded higher scores than massed practice. In particular, the GZLM analysis (Table 4.2) indicated that the type of practice had a highly significant effect on the participants' immediate post-test scores. Furthermore, there was a significant interaction effect of type of practice with vocabulary knowledge strength. The spaced practice scores were higher than massed practice scores by a different margin at different strength levels.

Additionally, results for scores based solely on vocabulary strength increased from the most demanding strength level to the least demanding strength level. However, further analysis (Table 4.3) indicated that the massed and spaced learning conditions differed significantly at only three strength levels (i.e., productive recall, receptive recall, and productive recognition) and barely missed significance at receptive recognition (p= 0.052). This lesser difference at the receptive recognition level was bound up with the fact that 77.6% and 65.4% of the participants scored 30 (out of a maximum of 30) in the spaced and massed learning conditions, respectively, and 20.4% and 24.5 % of the participants scored 29 (out of a maximum of 30) in the spaced and massed learning condition, respectively. As such, scores for receptive recognition were often 100% correct or nearly so, so ceiling effect was at work limiting the scope for any difference between learning conditions to show itself at the least demanding strength levels. Interestingly, the effect of learning condition in terms of difference in number of words learnt is at least twice as much at the more demanding strength levels than at the less demanding strength levels.

Consequently, the hypothesis of no massed-spaced difference in an immediate post-test, framed from the literature, was not supported, as massed practice and spaced practice produced similar initial vocabulary gains only at the receptive recognition level. These results depart from those of the study of Bloom & Shuell (1981) who found both massed and spaced practice equal in effect on receptive recall.

In interpreting the differences in results of the present study from the findings in Bloom and Shuell (1981), it is possible that some factors in Bloom and Shuell's (1981) study that are related to previous knowledge of the target items and L1 interference could explain these differences. First, Bloom and Shuell (1981) did not assess the participants' prior knowledge of the target vocabulary. If it was taken into consideration that the participants in their study were English speakers who were intermediate-level French learners, it is highly possible that some of the target vocabulary was already known by the participants which could have influenced their findings. Second, the participants in Bloom & Shuell (1981) were English-speaking learners of French, while in the present study the participants were Arabic-speaking learners of English whose L1 has few cognates with English and none among the words chosen in this study. It is quite possible that the degree of lexical similarity between English and French may play a role in how massed practice and spaced practice affect vocabulary learning, by making some words easy to learn regardless of the massed-spaced difference (Nagy, Garcia, Durgunoglu, and Hancin-Bhatt, 1992). In particular, the participants in Bloom & Shuell (1981) were learning French occupation names which are very common and have cognate similarities with English occupation names (e.g., French/English: chauffeur/chauffeur, électricien/electrician, pilote/pilot, journaliste/journalist). Therefore, if the words in the massed condition were more cognate than the words in the spaced condition, it is possible to conclude that the massed scores might have been artificially inflated. Consequently, this issue may explain why in Bloom and Shuell (1981), both massed and spaced practice had an equal in effect on the receptive recall level.

Note, however, that it was not possible to verify this interpretation since Bloom and Shuell (1981) did not publish the target words that were used in their study, but it is worth considering that interpreting and generalizing the findings of spacing effect studies should cautiously consider the degree of similarity between the L1 of the participants and the target language learnt. Thus, it could be argued that the findings in the present study describe the effect of massing and spacing of genuinely previously unknown English vocabulary among Arabic-speaking EFL learners.

6.2.2 The effect of repetition learning with or without spacing on long-term retention of vocabulary knowledge at four strength levels

RQ 2: What is the difference in effect of massed practice and spaced practice on the strengths of vocabulary knowledge acquired, as measured on a 4-week delayed post-test?

H2: It is expected that spaced practice leads to higher gains in vocabulary knowledge on delayed post-tests than massed practice. Research by Sobel et al. (2011) and Goossens et al. (2012) revealed better long-term recall of lexical items when learning opportunities were spaced over several learning sessions as opposed to massed in one single learning session.

Previous empirical research, reported in Section 2.6, suggested that spaced practice has an advantage over massed practice in long-term vocabulary retention. Therefore, it was expected in this study that the long-term retention of the target words taught with spaced practice would be significantly higher than that with massed practice. However, the answer to this question further revealed information unavailable from previous studies on the difference in effect of spaced practice and massed practice on long-term vocabulary learning at four degrees of strength of vocabulary knowledge.

Results of the delayed post-test suggested that vocabulary learning was in fact superior in the spaced condition to the massed condition across all four-strength levels of vocabulary knowledge. An overall GZLM-GEE analysis (Table 4.5) indicated that the effect of vocabulary strength was similar to that in the immediate post-test. However, the learning condition effect was much stronger in the delayed post-test. In addition, the comparison between spaced practice and massed practice at each strength level separately revealed that the two learning conditions differed highly significantly at every strength level and that the spaced scores were significantly higher than the massed scores at every strength level.

The scores increased successively across the four strength levels from the most demanding to the least demanding. However, while generally the delayed post-test results indicated bigger differences in vocabulary learning between spaced practice and massed practice than in the immediate post-test results, these differences were no longer more the case at both the more demanding strength levels but primarily just for receptive recall, which could be due to that level being the one that generated scores that were most distant from the floor and ceiling of the score scale. Hence, it had more 'space' on the score scale in which variation could emerge. Nevertheless, the higher differences found at all strength levels, in the delayed post-test compared to the immediate post-test, seem to demonstrate that spaced practice is especially effective in enhancing longer term retention, which is consistent with the wider massed-spaced literature.

6.2.3 The effect of repetition learning with or without spacing on attrition between immediate and delayed tests of vocabulary knowledge at four strength levels

RQ 3: Does massed or spaced practice yield better retention of vocabulary knowledge at any different strength level between the immediate post-test and the delayed post-test?

The relatively exploratory nature of the present study also provided new information and associated discussion regarding the effect of massed practice and spaced practice in terms of the difference between scores on the immediate post-test and the delayed post-test and the difference between the scores across the four degrees of vocabulary knowledge strength. In answering the previous two RQs, results for each of the two post-tests were examined separately, in relation to the zero-knowledge baseline before any intervention. However, in order to obtain the fullest picture of the effects of the two learning conditions on vocabulary learning, both occasions of post-testing were directly compared in one analysis, in relation to the knowledge strength levels and the two types of practice.

Results of an overall GZLM analysis (Table 4.7) suggested that there was a significant difference between the immediate and delayed post-test times, regardless of type of practice and strength, taking the form of a fall in scores between immediate and delayed post-tests. There was also a significant difference between types of practice regardless of time of testing and strength: spaced practice yielded higher scores than massed practice on both immediate and delayed post-tests at every strength level. In addition, there was a significant difference between vocabulary strength levels regardless of type of practice and time of testing, taking the form of a successive rise in scores from the most demanding strength level to the least demanding strength level in both immediate and delayed post-tests and in both learning conditions.

Furthermore, retention between immediate and delayed post-tests was not the same after both types of practice; types of practice had different effects at different strength levels regardless of time of testing; and differences between strength levels were not the same in both the immediate and delayed post-tests. Interestingly, the three-way interaction effect did not quite achieve significance (p = .096) which suggested that the learning condition by time of testing effect did not differ significantly between strength levels.

Further investigation revealed that scores for items that were learnt in the massed condition were not only lower than scores for items that were learnt in the spaced condition but also yielded a greater fall between the immediate and delayed post-tests, although that fall was not quite significant at the least demanding strength level (i.e., receptive recognition). It is essential to note that the scores were very low (floor effect) at the level of greatest difficulty of vocabulary knowledge (i.e., productive recall) and very high (ceiling effect) at the level of least difficulty of vocabulary knowledge (i.e., receptive recognition); these two effects of the closed ends of the score scale made differences in the loss of vocabulary knowledge due to learning condition harder to distinguish at those levels of knowledge. For the other two levels where scores were more distant from the extremes of the scale, vocabulary loss was considerably greater at the receptive recall level than at the productive recognition level.

Although the effect of massed and spaced practice was difficult to identify at the least and most demanding levels, the findings further support the notion that spaced practice yields better vocabulary retention than massed practice.

6.2.4 The spacing effect and parts of speech

RQ 4: Do word classes differ in lending themselves to being initially learnt and/or retained at different strength levels if practiced spaced or massed?

The effect of massed practice and spaced practice on learning EFL words of different word classes has not been explored in previous research. Only one study to date examined the effect of massed and spaced practice in relation to word class but that concerned L1 vocabulary learning among two-year-old children (Childers & Tomasello, 2002).

Results in the current study revealed that there was a significant difference overall between scores for nouns and verbs, with the former higher regardless of the time when the test occurred (Table 4.12). Furthermore, results also indicated that there was a significant interaction effect between time of testing and word class: not only did verbs score lower than nouns, but scores for verbs also dropped between immediate and delayed post-tests further than those for nouns. Additionally, there was a significant interaction effect of word class with strength: scores for nouns were higher than scores for verbs by a slightly greater margin where recall rather than recognition was tested.

Results crucially also revealed, however, that the word class by learning condition interactive effect was far from significant. This finding indicated that the benefit of spaced practice over massed practice applied equally to verbs and nouns. Given that verbs seem to be at a disadvantage in other ways compared with nouns (e.g., in initial learning and retention) it is encouraging to find that their learning was not affected differentially worse by the type of spacing used. On the other hand, it was not the case either that spaced practice significantly reduced the disadvantage of verbs in relation to nouns.

Finally, the three-way interaction effects were only significant for time by word class by strength: scores for verbs fall more than nouns between immediate and delayed post-tests at different strength levels. However, it was noticeable that the fall of scores from the immediate post-test to the delayed post-test was the same for nouns and verbs at the most demanding level, that is, productive recall. A possible reason for this finding was again however the purely technical one of floor effect which especially affected verbs.

6.2.5 Learners' perspectives on benefits for learning and engagement of the spacing effect

RQ 5a: How far do participants perceive any difference in the benefits of spaced or massed vocabulary learning?

It is essential to note that earlier attempts to examine the difference in effect of massed and spaced practice on EFL classroom vocabulary learning did not assess learners' perspectives on the benefit of massed and spaced practice for learning and engagement (e.g., Sobel et al., 2011; Goossens et al, 2012). However, some wider cognitive psychology research suggested that learners do not perceive any difference in learning between massed and spaced practice, despite exhibiting higher memory performance in spaced practice relative to massed practice (e.g., Logan, et al., 2012). The answer to this question is divided into two parts reflecting learners' perceived benefit for learning and learners' perceived benefit for engagement (i.e., motivation and attention) of the spacing condition.

Regarding learners’ perceived benefit for learning of spaced over massed practice, many students remained neutral and did not indicate a preference for either type of practice. Of those who did express a preference, the majority descriptively preferred massed practice. There was a significant preference, however, only on one item (i.e., *I memorize more words in spaced learning than massed learning*), where participants agreed that they thought massed practice led to more memorization. These results did not correspond to the participants' actual performance on the post-tests since their scores were higher for items learnt in the spaced condition than items learnt in the massed condition. A possible explanation for these results could be that the participants did not judge correctly where they learn more or less successfully, something which was found in other studies (Dunlosky & Nelson, 1994; Simon & Bjork, 2001; Kornell & Bjork, 2008; Kornell et al. 2010; Logan et al., 2012; Bjork, et al., 2013).

Regarding learners’ perceived engagement with spaced over massed practice, again many students remained neutral. Among those who expressed an opinion, however, results revealed a highly significant majority who thought massed practice was better in terms of being able to focus and causing less boredom. It is understandable that massed practice seemed more appealing to learners in terms of being able to focus on learning than spaced practice since it delivered information in one extended session rather than in fragments over different sessions. For the same reason, it made sense that learners felt bored while learning vocabulary with spaced practice as it involved revisiting the same words on many different occasions.

6.2.6 The effect of learners’ perceived benefit for learning and engagement on the spacing effect

RQ5b: What is the relationship between a learner's degree of favouring of massed or spaced learning and their retention of vocabulary knowledge of each strength between the post-test and 4-week delayed post-test?

In order to answer this question, several correlation analyses were performed. The analyses included the possible relationship of the participants' questionnaire ratings with immediate and delayed post-test scores at each vocabulary strength level, and with degree of change of scores between immediate and delayed post-tests. The results of these analyses suggested, however, that in no instance was any correlation significant.

Consequently, it would appear that there was no relationship between the learners' perceptions of benefits for learning, or their reported engagement with massed and spaced practice, and how well the participants performed with them. Accordingly, this result implied that vocabulary learning with spaced practice was beneficial to all learners irrespectively of whether they preferred it or not over massed practice.

6.3 Limitations of the study

Despite careful preparation of the current study, there were some limitations which were not possible to avoid in some cases. It should be noted, however, that despite the limitations of the study, it is still methodologically stronger than previous studies.

One of the limitations in this study was participant attrition. From the start of the teaching phase to the delayed post-tests, the number of students who participated in all the sessions of the study dropped from 62 to 49 participants. However, this issue is unavoidable in studies that involve longitudinal data from human subjects.

Another apparent limitation concerns the fact that all participants in the present study were male Taif University students. Therefore, the limited representativeness of participants would suggest some caution before generalizing the findings of the study. The selection of gender was dictated by the religious restrictions of the context which impacts on research in Saudi Arabia. However, there should be no reason from other studies to suppose that there is any gender difference with respect to the massed-spaced effect.

The choice of Taif University students was motivated by convenience. Since I hold the position of lecturer at the university, it was easy to obtain permission to personally resume formal teaching of an EFL course in an actual classroom setting for three months, as was required by the design of the study. This would not have been possible to organise at any other university. Furthermore, aside from the general interest of the project for the EFL vocabulary research community, I was motivated in part by the action research aim to help my own students to improve their learning. That again necessitated that I focus on students that I would normally teach in my own university. However, there is no reason to suppose that Taif University students differ markedly from students at other public universities in the KSA. Nor indeed is there any obvious reason why their vocabulary learning behaviour would differ from that of tertiary level EFL learners in many countries around the world.

As explained in detail in Chapter 3, random selection of participants was not possible because using intact classes was an inherent feature of the study (i.e., the study was embedded into an actual EFL course). The design of this study may have arguably reduced the internal validity of the study, but it increased its ecological validity which is necessary where, as I did, a researcher wants the study to best resemble actual educational settings as compared to controlled laboratory settings (Gu, 2003). It is common in educational research for there to be some compromise made between experimental control of unwanted factors, which makes for a rather artificial context, and less control of such factors, in order to maintain a reasonably natural context. In this study, a natural context was chosen so as to be able to use the findings to make recommendations about teaching in real situations in the KSA. Furthermore, since with respect to the two learning conditions, a repeated measure rather than a between groups design was used, it should be noted that the use of intact classes cannot have biased the results in favour of either treatment. The participants were their own controls, in the sense that any relevant individual differences that they possessed were the same for learning in both the massed and spaced conditions.

Another limitation on the generalisability of the results could arise from the fact that the target words in this study were selected from only two word classes (i.e., 30 nouns, 30 verbs). Hence, no claims can be made about the effect of word class beyond those two parts of speech. According to Schmitt (2010), it is necessary to control for word class in all vocabulary research by either using the same number of items for each word class or using only one type of word class. However, including the same number of items for more than two types of word class in this study (i.e., including adjectives and adverbs as well as nouns and verbs) would have resulted in increasing the amount of the target words to 120 words and extending the treatment time for massed practice to an uninterrupted three-hour session, which would take all the allocated three hours lecture time. On the other hand, if all four main types of word class (i.e., noun, verb, adjective, adverb) were to be included while maintaining the same overall treatment time for massed practice as there was in the current study, only six or seven words from each word class would be practiced in each learning condition which might be insufficient to draw any valid conclusions.

In addition, it must be admitted that some of the target words, which for the study were only taught and tested in one part of speech each, can in fact function as both nouns and verbs (e.g., *defeat*, *object*, *escape*, *permit*, and *harvest*). This limitation might not be an issue however since the pre-test had an option for the participants to list other possible meanings or translations for each target word. Furthermore, since the participants did not indicate any pre-existing knowledge of the target words, it is possible to claim that they did not know the target words in any part of speech prior to the intervention, so only learned new words in the part of speech which was chosen for the study and delivered in the presentation, practice and tests.

It is however possible to argue against the perfect validity of a pre-test such as the one that was used in this study as a means of detecting prior knowledge of aspects of words including meaning and part of speech. Although such pre-tests are widely used, it may be that participants cannot easily recognize and externalize their own vocabulary knowledge or perhaps may give no indication due to lack of confidence in the accuracy of their knowledge. For example, in Schmitt (1998) the participants had to be probed deeply to talk about all of their vocabulary knowledge. As such, it is possible in the present study that even with an option in the pre-test for the participants to state whether they know other meanings/ translations of the target words, some participants could have been familiar with the forms, part of speech or even meanings of some of the words yet still not revealed that. Hence, later learning for some words may have actually involved learning a new meaning and / or part of speech for a known form.

Furthermore, the current study assessed the effect of massed and spaced practice on vocabulary knowledge of form and meaning at four degrees of strength (i.e., receptive recognition, productive recognition, receptive recall, productive recall). While this is more than many studies do (indeed none have done it in the massed-spaced literature prior to this study), it does not of course embrace all the kinds of word knowledge that exist and could be tested. For instance, knowledge of lexical collocation was not covered, such as the fact that a *thread* would often be mentioned in the same context as a *needle* and the verb *sew*. Similarly, knowledge of the grammatical idiosyncrasies of the words was not tested, such as the irregular past tense form of *shrink* or the fact that one *forbids* someone *from* doing something (complement type, and preposition choice). Student knowledge of register and stylistic values was not tested either, such as that *trauma* is a medical word and *comply* somewhat formal. This sort of information is more relevant to production than reception use of words, since the speaker or writer has to supply it in L2 in production but only decode it in reception. It is possible that some of this information was in fact learned, through the example sentences used in the retrieval practice material. However, it is not practically possible to assess every aspect of vocabulary knowledge at once and assessing vocabulary knowledge of form and meaning in the present study was rational since knowledge of form and meaning is considered the basic knowledge in learning vocabulary (Schmitt, 2010).

A ceiling effect was observed at the least demanding strength levels and a floor effect at the most demanding strength level (ceiling=30, floor=0). However, ceiling and floor effects were also reported in other studies examining the spacing effect, such as Childers and Tomasello (2002), Seabrook et al., (2005), Ambridge et al., (2006), Rohrer (2009), and Walsh et al., (2018), which suggests that ceiling and floor effects in the study of massed and spaced practice are difficult to avoid. Arguably, it might be possible to assume that a ceiling effect could be avoided by including more words in the study, so that it becomes hard for many participants to achieve scores that approach the maximum, even at the least demanding knowledge level (i.e., receptive recognition). However, the same cannot be so easily done to avoid floor effect, which was present for the most demanding level (i.e., productive recall), since reducing the number of words to be learned in order to make the task easier would damage the reliability of the study. Thus, as a limitation, ceiling and floor effects remain difficult to avoid in vocabulary studies that aim at examining the effectiveness of different learning methods, especially an effective learning method such as spaced practice.

One final apparent limitation was the possible practice effect from one test to another since all tests, at different times and at different levels of knowledge strength, assessed knowledge of the same target words. However, in this study, test repetition at the different times (i.e., pre-test, immediate post-test, and delayed post-test) is less of an issue because considerable time elapsed between those occasions, and this is a common feature of repeated measure designs. More crucially, however, the very nature of vocabulary strength/depth tests involves testing the same target items at different knowledge strength levels close together on each testing occasion. Generally, the internal validity of studies may be threatened by practice effects if the items in the tests are presented in the same order every time to all participants very close together in time. In order to mitigate the practice effect of vocabulary strength tests, therefore, Laufer et al. (2004), have suggested testing two strength levels in one sitting and another two strength levels at another sitting with a delay gap of around one week to minimize practice effect. However, since the aims of Laufer et al. (2004) and the aims of this study are different, this method was neither suitable nor practical. For example, if two parts of the immediate post-test were administered immediately after the end of the practice treatments and the other two parts a week later, this would not have allowed a fair comparison between the effect of massed and spaced practice on initial gains of target words knowledge on all four strength levels. Consequently, as far as possible, this study used a number of methods other than time distancing to diminish the practice effect between the sub-tests. First, the order of each item in each subtest in the immediate and delayed post-tests was randomly re-ordered. Second, in multiple-choice subtests (i.e., respective recognition and productive recognition tests) in the delayed post-test, the distractors were randomly reallocated to each target item while, at the same time, word class and word length were maintained consistent between the target words and distractors.

It should be noted that earlier spacing effect classroom-based studies often implemented a research design with only delayed post-tests to compare between the effect of massed and spaced practice (e.g., Sobel et al., 2011; Goossens et al., 2012). The current study adopted an immediate post-test and delayed post-test design which should be factored into the interpretation of results because the immediate post-test in effect provided additional exposures to the target items (i.e., another set of four repetitions), with very short intervals between them, beyond those in the practice sessions. In this way the test could have had some instructional/practice benefit and so could have led to better scores on the delayed post-test than would have been obtained if there had been no immediate post-test. However, the current study adopted the recommendation of a minimum delay of four-weeks between the immediate post-test and delayed post-test that according to Schmitt (2011, p. 157) should be “indicative of learning which is stable and durable”. Nevertheless, it could be prudent to cautiously assume that the results obtained in this study are only fully comparable to those from other studies that adopted a similar research design (i.e., immediate post-test-delayed post-test with delay length of four-weeks between tests).

6.4 Implications for SLA and vocabulary learning theory

The nature of vocabulary acquisition is not fully understood in the SLA field of vocabulary research (Nation, 2001; Schmitt, 2008; Staehr, 2009). A number of themes explored in the literature review (see Chapter 2), which this study has implications for, are considered here.

6.4.1 Productive and receptive vocabulary knowledge

In general, it is agreed among researchers that vocabulary learning is incremental in nature and that some aspects of vocabulary knowledge are acquired before others (Gass, 1999; Schmitt, 2010). However, there are inconsistent accounts in the SLA literature on the order of acquisition of some aspects of vocabulary knowledge (Laufer & Goldstein, 2004). In general, it is claimed or found that productive vocabulary knowledge lags behind receptive vocabulary knowledge (Laufer & Goldstein, 2004), as words tend to be initially known receptively then over time become known productively (Schmitt, 2010). However, it is also claimed that vocabulary knowledge could start from productive knowledge, where learners are immediately expected to say aloud and write new words, only after which the words are encountered receptively (Milton, 2009).

An overwhelming support for receptive knowledge being better known than productive knowledge was found in the current study, based not only on the mean scores but also on an implicational scale analysis of individual scores of participants across strength levels, considered for each word in individual tests. This support can arguably be interpreted as a good evidence for order of acquisition of these types of knowledge, although it should be admitted that more thoroughgoing longitudinal studies of individual cases are needed to provide absolutely conclusive proof.

Furthermore, the practice exercises in the current study were done by participants in that same order (receptive knowledge exercises before productive knowledge ones), so arguably the participants may have been primed for the standardly assumed scenario referred to by Schmitt

(2010) rather than that of Milton (2009). Again, it would require a separate study to determine whether, regardless of the order in which the students do the exercises, they still end up getting better scores for receptive than productive knowledge.

Results in this study also threw important light on why a consistent result may not always be obtained for the primacy of receptive and productive knowledge. This relates to the test item mode through which vocabulary knowledge measurement is obtained. As indicated in Chapter 2, Section 2.10, productive knowledge is often tested in open response items (e.g., PVLT, Laufer & Nation, 1995, 1999) while receptive knowledge is often tested in multiple choice mode (e.g., VLT, Schmitt, et al., 2001), but this confuses the test item support mode (i.e., open response recall versus multiple choice recognition) with the receptive-productive knowledge distinction. Hence, the present study carefully distinguished those two oppositions and tested all the four possible combinations, recognising that there are degrees of strength within receptive vocabulary knowledge (i.e., receptive recognition, receptive recall) and within productive vocabulary knowledge (i.e., productive recognition, productive recall) which could shed light on the issue of the inconsistency in the literature with respect to the primacy of receptive or productive vocabulary knowledge.

Indeed, results in the current study indicated that the recall-recognition dimension had a far greater impact on vocabulary test scores than the receptive-productive dimension. This finding therefore raises some questions about the difference between receptive and productive vocabulary knowledge in previous studies (e.g., Waring, 1997; Laufer, 1998; Laufer & Paribakht, 1998; Fan, 2000). One might argue that the difference between receptive and productive vocabulary knowledge has been wrongly exaggerated by confusing receptive knowledge with recognition and productive knowledge with recall. It can be seen clearly from the findings in the present study that if a study tests receptive vocabulary knowledge only in recognition mode (e.g., multiple-choice questions) and productive knowledge only in recall mode (e.g., open response) then the difference will look much bigger than if both receptive and productive vocabulary knowledge were tested in recognition mode.

Consequently, it is worth pointing out that any examination of differences between receptive and productive knowledge should seriously consider the effect of what response mode is selected for each. For instance, assessing knowledge in the form of receptive recall and productive recognition could, by reason of the mode rather than the productive-receptive difference, produce a result that made reception and production appear quite similar in difficulty. I would therefore urge a cautious approach to this issue due to limited existing research on strength/depth of vocabulary knowledge.

6.4.2 Vocabulary learning rate

Although not its main purpose, the present study provides some evidence relevant to understanding the optimum rate of learning new words, and hence the rate at which teachers and materials should present them. This again is a topic riddled with different opinions and estimates. For example, Scholfield (1991) suggested a rate of nine words per hour for a hypothetical course, based on 'rule of thumb' suggestions to be found in the literature at the time. Figure 1 shows how this might be distributed over units in a hypothetical course which in fact contained revision sessions where no new vocabulary was introduced but previously learned vocabulary was recycled, or indeed made to occur in spaced practice exercises.

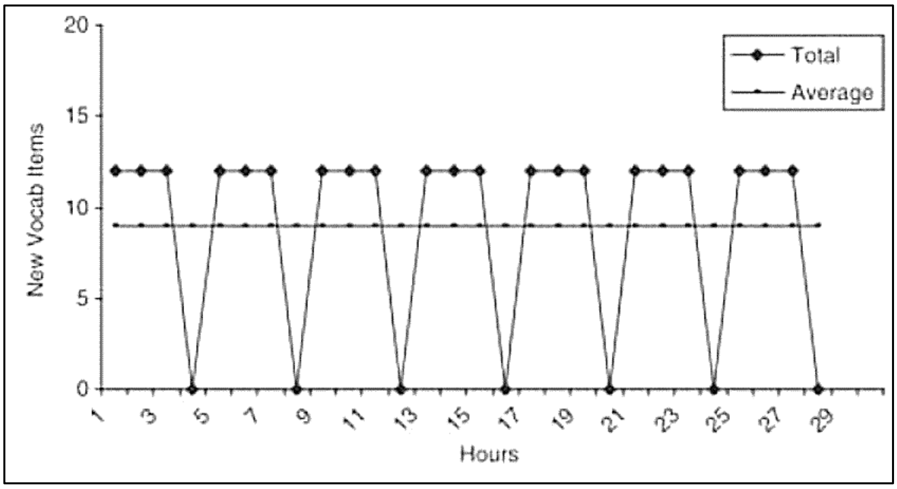


Figure 6.1 Vocabulary teaching rate suggested for a theoretical course (Scholfield, 1991, p. 27)

However, by looking closely at the rate of learning per hour in the present study, it would seem that specifying a general average rate of vocabulary learning is not simple since different assessment tools testing different knowledge strength levels will suggest different outcomes. For example, based on the immediate post-test results in the present study for the spaced practice, the participants initially learned an average of 29 words per hour at the receptive recognition level, 27 words per hour at the productive recognition level, 15 words per hour at the receptive recall level, and only 6 words per hour at the productive recall level. Accordingly, the hourly rate of learning words at different vocabulary knowledge strengths seems to vary, and any suggestion in regard to how many new words per hour/class a textbook, class lecture, or syllabus should introduce must be made in relation to the target level of vocabulary knowledge.

The study's findings of course mostly suggest much higher learning rates than those referred to by Scholfield (1991). That, however, is influenced by the fact that an hour in the current study was an hour of nothing but vocabulary learning, while Scholfield was thinking of a vocabulary learning rate in a normal course where vocabulary was not the sole topic of an hour of lesson time. That therefore indicates another important consideration in estimating any vocabulary teaching/ learning rate: what proportion of lesson time is imagined to be devoted to vocabulary.

6.4.3 Part of speech effect on vocabulary learning

The result that was obtained for learning of nouns versus verbs replicates the typical finding in the literature, which indicates that nouns are easier to learn than verbs (e.g., Laufer, 1997; Horst & Meara, 1999; Al Fotais, 2012). However, the current study is the first to my knowledge to show that the massed/spaced practice difference does not differentially affect learning of words of different word classes. For comparison, some vocabulary learning techniques, which, although effective on some words, are clearly not suited to certain parts of speech. One such method of vocabulary learning is Total Physical Response, which struggles with many nouns (Oxford and Crookall, 1990).

6.4.4 Number of repetitions and noticing

It has been noted for decades (see Section 2.3.5) that spaced repetition over an extended period of time (not just rote repetition on one occasion) is necessary for consolidating and maintaining vocabulary knowledge. There is a long history of recommendations, summarised in works such as Nation (2012), which focus on ways of ensuring that new vocabulary items are met repeatedly over an extended period, often under the name of recycling.

As discussed in Section 2.6.5, the effect of vocabulary textbook recycling in particular has been linked to increasing the chance of learning vocabulary (e.g., Vassiliu, 2001; Alsaif, 2011) and, for example, there are studies showing that at least five and maybe fifteen recycled encounters may be needed to learn a word (Nation, 1990; Rott, 1999; Webb, 2007). The number of repetitions in the previous studies seem more than the number in the current study, or in massed-spaced studies generally. On closer inspection, however, it may not be that much greater if one count as repetitions the fact that each word occurs at least twice in each exercise, once when the student does the exercise and again a little later when the teacher gives the correct answers, and that the immediate post-test provided four extra exposures to each word. Together, and not including initial presentation, that adds to 12 occurrences before arriving at the delayed post-test.

In any event, simple re-encounter with a word, whether in a textbook or reading materials, often occurs however incidentally and implicitly, in such a way that the focus is not on the word but on something else, such as the message of a text, or maybe on grammar if a word recurs in a sentence used in a grammar exercise. In accordance with much of the SLA research (e.g., Schmidt, 1990,1992; Laufer, 2006; Nation 2013), it would be expected that spaced learning through such recycling would be far less effective, and require more repetitions, than repetition in the form of explicit and intentional vocabulary exercises on words met before, as in the current study, since in the latter case the learner is forced to notice the lexical information.

The kind of spaced-massed research paradigm that was adopted in the present study in fact has always considered only fully noticed repetitions, using words like *study* or *review*, which imply conscious attention, to characterise what occurs at each repetition. Hence this study complements work on how many the not-necessarily-noticed repetitions in recycling research need to be. Results of the delayed post-test indicated that it was possible to achieve a 97% learning success (29 out of 30 maximum score) at the lowest strength level with spaced repetition of 16 noticed exposures (i.e., initial presentation, four exercises, four corrective feedback, four immediate post-tests, and three delayed post-tests). Even with massed repetition, the learning success was 91% (27 out of 30 maximum score) at the same strength level and with the same number of noticed exposures.

Receptive recognition represents perhaps the level of word knowledge which is most often in effect referred to in studies of learning through recycling. Compared with estimates of the number of repetitions required where largely unconscious learning through recycling occurs, the figures in the current study show a considerable advantage in the number of noticed repetitions needed to achieve almost 100% acquisition through spaced learning. Nevertheless, it should be admitted that at higher levels of knowledge, spaced learning only achieved quite a low success rate: productive recall immediate post-test 21%, delayed post-test 14%.

6.4.5 Spacing of repetitions

All vocabulary learning which involves more than one encounter with a word may be analysed in terms of the spacing between the repetitions. Massed learning is based simply on repetitions with zero time between them. Anything else is spaced but can differ in the lengths of time that create the spacing, so this aspect of spacing has attracted attention from researchers, perhaps rather more in psychology than SLA (see Chapter 2, Section 2.8). The range of spacing varies, however. Donovan and Radosevich (1999) record some studies with spaced conditions using intervals as little as one minute, while others had intervals of many days.

Unconscious learning when words are recycled often involves repetitions with uniform, increasingly varied, or randomly varied spacing, while the kind of spaced conscious learning which has been mainly considered in this study typically has involved uniform periods of time between each repetition. In the present study and Sobel et al (2011), the spaced practice for vocabulary learning was based on one-week gaps between repetitions, while a similar result was found with only a one-day gap between repetitions in Goossens et al. (2012). Thus, the present study has confirmed the effectiveness of uniform gaps of one week.

The optimum spacing clearly remains uncertain. Donovan and Radosevich (1999) found surprisingly that the effect size of spaced over massed learning in a range of studies actually became systematically higher as the interval between repetitions became shorter. However, they put this down to the fact that many of the tasks in studies they reviewed were purely motor tasks (unlike the current study), indicating that the best interval has to be determined for each type of task separately.

6.4.6 Student attitudes to massed/spaced learning

It is common in language learning and teaching research to find that students have positive attitudes to methods which are less effective, or negative attitudes to those that are more effective (Garrett et al., 1994). In the case of the participants in this study, where they expressed a preference at all, they predominantly preferred massed practice which was objectively less effective.

6.5 Implications for EFL pedagogy in Saudi Arabia and more generally

The present study examined the effect of massed and spaced practice on vocabulary learning in actual EFL classrooms. The findings revealed that spaced practice led to better gains and retention of vocabulary knowledge than massed practice. Most importantly, spaced practice was found equally effective for different word classes (although nouns remained, as usual, easier overall to learn than verbs), and to be beneficial regardless of the learners' preference/perceptions with respect to massed or spaced practice. Due to the important role that vocabulary learning plays in language learning, teachers and materials designers should seek the most effective way of teaching vocabulary to language learners. Therefore, in light of the study's results it seems logical to propose the implementation of spaced conscious practice in classroom vocabulary learning. What remains to be decided is how exactly to achieve this, and whether the key agents should be the textbooks or other resources (e.g., computers), the teacher, or the students themselves.

Only a few empirical studies have been conducted to examine the spacing effect in authentic EFL classroom environments and they were in very different educational contexts to the current study, both in terms of L1 and age (e.g., Sobel et al., 2011; Goossens et al., 2012). Therefore, one can only offer the following suggestions tentatively.

Brown (2011, p. 93) suggested that an effective textbook “introduces a set of items and then regularly revisits those items by looking at other aspects of knowledge”. As such, it is advisable to recycle vocabulary in a principled way in subsequent activities that focus on a different aspect of vocabulary knowledge. Recycling in the broad sense (Nation 2012) may be variously engineered by textbooks or graded readers making sure that new items do not only appear in one unit or chapter of the book, but in successive units later. Furthermore, recycling may also be ensured by teachers who organise quizzes on vocabulary that was introduced a week, a month or even several months previous. It may even be organised by students themselves keeping a vocabulary notebook or vocabulary cards or a computerised version of that and testing themselves on vocabulary from longer distances in the past rather than just one or two days (e.g., Sheridan and Markslag, 2017).

Based on findings in the current study, it seems sensible to further suggest that such longer-term repetition of vocabulary should also occur in the form of classroom vocabulary exercises that are introduced at systematic intervals. That is a step beyond what recycling is usually understood to mean, which is typically only the re-encountering, or at most re-testing, of the same items, and not following any fixed pattern of gaps between those events.

Spaced practice can be practically utilized by teachers and language instructors in many educational settings. For example, Goossens et al. (2012) used three repetitions on successive days. Such spaced practice can be easily applied to EFL classrooms in Saudi public schools since English language courses are usually taught in one 45 minutes class per day for four days in the week. Moreover, if one considers the same spacing schedule that was employed in the present study (i.e., four repetitions in successive weeks), then spaced practice can be easily applied to tertiary level education since EFL courses/lectures usually occur once every week. As mentioned above, only a few empirical studies have been conducted to examine the spacing effect in authentic EFL classroom environments so it is difficult to suggest a minimum number of spaced repetitions or a minimum period of delay between each repetition. However, the key notion here is that spaced practice is more efficient in promoting initial gains and retention of vocabulary knowledge than massed practice (or indeed no practice!).

A crucial point is that the practice needs to be vocabulary focused, so as to ensure noticing, and involve exercises in a fixed time schedule, as distinct from haphazard recycling of word encounters where the learner's attention may be directed to something other than vocabulary. Specifically, materials writers need to consider spaced repetition of exercises as an important factor when designing language learning materials. This step of incorporating systematically spaced exercises on the same words, such as in the present study, across a range of units is not however typically found at present.

It should be noted that there is evidence that any kind of vocabulary recycling is a neglected aspect in many EFL textbooks (e.g., Nation, 1990; Fujimori, 2005; Al Fotais, 2012), let alone the introduction of spaced exercises. For example, Al Fotais (2012) examined two vocabulary textbooks taught at university level in Saudi Arabia. The results revealed that around 81% of the vocabulary in the textbooks was not recycled beyond its initial presentation.

It could be argued that teachers may overcome the issue of limited vocabulary recycling in textbooks by themselves explicitly recycling vocabulary after initial learning from textbooks. However, some studies suggested that teachers may tend to be selective and recycle only frequent words that would assist their students to understand them in the classroom (Meara et al., 1997; Tang & Nesi, 2003). Based on the results in this study, I would suggest that vocabulary learning exercises should occur spaced in university textbooks at least three times after initial presentation of each target word.

A final possible suggestion is concerned with the type of spaced or recycling activities. It is generally agreed that textbooks should recycle vocabulary using interesting and refreshing activities (Harwood, 2002), which ideally provide learners with further knowledge about different aspects of the target words (Brown, 2011). In the present study, the impact of different learning activities or tasks was not examined beyond the range necessary to practice the four levels of lexical knowledge that the study targeted. Still, the high levels of learning of some types of lexical knowledge in both learning conditions may be due in part to the range of practice tasks used, targeting knowledge of form and meaning of each target item in four different ways. Boredom might well have ensued if the same kind of vocabulary exercise had been used four times.

6.6 Future research directions

As mentioned earlier, the spacing effect is one of the most robust findings in memory research and is relevant to the educational context. Nevertheless, few studies have examined the spacing effect using vocabulary practice exercises in authentic EFL vocabulary learning classrooms, as distinct from randomly spaced recycling of vocabulary encounters, typically without learner attention necessarily being consciously focused on the words. Conclusions from the present study therefore can act as a starting point to motivate more investigations, or as models for other researchers to validate in follow up research.

First, it is noteworthy that different studies on the spacing effect utilized different methods and procedures in the study of the spacing effect in EFL classrooms (see Chapter 2, Section 2.9) but, for whatever reason (e.g., the space exigencies of journal articles), these studies did not always make clear exactly what they did. The lack of clear and detailed methodological accounts in the previous studies makes comparing or following up on results of these studies very difficult. Therefore, the current study avoids this issue by being as explicit as possible (see Chapter 3). In any future work, however, a detailed report of the methods of massing and spacing used by empirical studies is a prerequisite for the development of this field of research and, consequently, for valid implications to be drawn from the research which may lead to improvements in pedagogy.

Having said that, a number of areas do seem to call for further attention. The majority of classroom spacing studies on EFL vocabulary learning, the present one included, examined the effect of spacing on learning single words. Therefore, it would be recommended that future investigations examine some of the larger units of vocabulary, such as collocations and phrasal idioms. Furthermore, as indicated earlier, there exist many areas of vocabulary knowledge to be learned beyond the basic single form and meaning, considered in two directions (reception and production) and with two types of support (recognition and recall). Grammatical knowledge that is specific to words, stylistic, registral and connotational/associative information about words, as well as multiple meanings (polysemy) of a single word all wait to be investigated in massed versus spaced practice studies. The current study adopted Laufer and Goldstein's (2004) description of vocabulary knowledge and examined the spacing effect at four vocabulary strength levels of form and meaning (i.e., receptive recognition, productive recognition, receptive recall, and productive recall). It would be interesting however to examine whether the benefit of the spacing effect extends beyond learning such basic knowledge of vocabulary in the classroom.

Furthermore, future research could also investigate the effect of massed and spaced practice on word classes other than nouns and verbs. It might be reasonable to assume that other word classes equally benefit from spaced practice. However, the examination would further shed much needed light on the effect of spaced and massed practice on learning different parts of speech.

Finally, there is much to be found out by varying the time gaps between repetitions in the spaced condition from those used in the present study, including examining the effect of successively increasing gaps. Given the shortage of massed-spaced studies of learning EFL vocabulary, there is at present no certainty as to whether shortening or lengthening the spacing between the classroom exercises, or systematically increasing it, would have any differential effect on the advantage of spaced over massed instruction.

Furthermore, the effect of time on long-term vocabulary learning in the current study was clearly present in the form of vocabulary attrition between the immediate post-test and the delayed post-test in both learning conditions. The delayed post-test in the current study was administered after four weeks from the end of the treatment/immediate post-test which is in line with the recommended minimum delay period for delayed post-tests (Schmitt, 2010). To further capture the long-term durability of vocabulary learning with spaced and massed practice, it would be suggested to increase the delay period after administering the immediate post-test to six weeks or more.

Bibliography

Abdan, A. A. (1991). An exploratory study of teaching English in the Saudi elementary public schools. *System, 19*(3), 253-266.

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons, Inc.

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108-119). Tokyo: Taishukan Publishing Company.

Al-Akloby, S. (2001). *Teaching and learning English vocabulary in Saudi Arabian public schools*. (Unpublished doctoral dissertation). University of Essex, Colchester, UK.

Alenezi, A. (2012). *Faculty members' perception of e-learning in higher education in the Kingdom of Saudi Arabia (KSA).* (Unpublished master's dissertation). University of Essex, Colchester, UK.

Alenezi, S. (2016). *The suitability of the EFL reading texts at the secondary and preparatory levels as a preparation for academic reading at first year university level in Saudi Arabia.* (Unpublished doctoral dissertation). University of Essex, Colchester, UK.

Alfotais, A. (2012). *Investigating Textbooks Input as a Possible Factor Contributing to Vocabulary Knowledge Failure among Saudi EFL Learners at Taif University*. (Unpublished master's dissertation). University of Essex, Colchester, UK.

Alhaidari, M. (2006). The effectiveness of using cooperative learning to promote reading comprehension, vocabulary, and fluency achievement scores of male fourth-and fifth-grade students in a Saudi Arabian school. (Unpublished doctoral dissertation). Pennsylvania State University, USA.

Alhaisoni, E. (2012). Language learning strategy use of Saudi EFL students in an intensive English learning context. *Asian Social Science, 8*(13), 115.

Alhatmi, S. A. (2012). *An investigation into the use of the vocabulary note-taking strategy by university EFL learners in Saudi Arabia.* (Unpublished doctoral dissertation). University of Essex, Colchester, UK.

Al-Hazemi, H. A. A.-G. (1993). *Low-level EFL vocabulary tests for Arabic speakers.* (Unpublished doctoral dissertation). Swansea University, Swansea, UK.

Al-Hazmi, S. (2003). EFL teacher preparation programs in Saudi Arabia: Trends and challenges. *Tesol Quarterly, 37*(2), 341-344.

Al-Hazmi, S., & Scholfield, P. (2007). Enforced revision with checklist and peer feedback in EFL writing: The example of Saudi university students. *Scientific Journal of King Faisal University, 8*(2), 237-267.

Aljasser, F. M., Jackson, K. T., Vitevitch, M. S., & Sereno, J. A. (2018). The influence of phoneme inventory on elicited speech errors in Arabic speakers of English. *The Mental Lexicon, 13*(1), 26-37.

Al-Nujaidi, A. H. (2003). *The Relationship Between Vocabulary Size, Reading Strategies, and Reading Comprehension of EFL Learners in Saudi Arabia.* (Unpublished doctoral dissertation). Oklahoma State University, Oklahoma, USA.

Alrabai, F. (2014). A Model of Foreign Language Anxiety in the Saudi EFL Context. *English language teaching, 7*(7), 82-101.

Alsaif, A. (2011). *Investigating vocabulary input and explaining vocabulary uptake among EFL learners in Saudi Arabia.* (Unpublished doctoral dissertation). Swansea University, Swansea, UK.

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal, 40*(1), 21-33.

Al-Masrai, A., & Milton, J. (2012). The Vocabulary Knowledge of University Students in Saudi Arabia. *Perspectives (TESOL Arabia), 19*(3).

Al-Masrai, A., & Milton, J. (2015a). Word Difficulty and Learning among Native Arabic Learners of EFL. *English language teaching, 8*(6), 1-10.

Al-Masrai, A., & Milton, J. (2015b). An investigation of the relationship between L1 lexical translation equivalence and L2 vocabulary acquisition. *International Journal of English Linguistics, 5*(2), 1.

Al-Seghayer, K. (2015). Salient Key Features of Actual English Instructional Practices in Saudi Arabia. *English language teaching, 8*(6), 89-99.

Al-Shuwairekh, S. (2001). *Vocabulary learning strategies used by AFL (Arabic as a foreign language) learners in Saudi Arabia.* (Unpublished doctoral dissertation). University of Leeds, Leeds, UK.

Alyami, S. (2011). *Vocabulary learning strategies of Saudi EFL majors of different gender, year, and proficiency: use and reasons for use.* (Unpublished doctoral dissertation). University of Essex, Colchester, UK.

Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive development, 21(2), 174-193.*

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

Andersen, E. B. (1997). The rating scale model. In van der Linden, W. J., & Hambleton, R. K. (Eds). *Handbook of modern item response theory* (pp. 67-84). New York, NY: Springer.

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

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296.

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*(4), 566-577.

Bahrick, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(2), 344.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and aging, 4*(1), 3.

Bao, G. (2015). Task type effects on English as a Foreign Language learners' acquisition of receptive and productive vocabulary knowledge. *System, 53*, 84-95.

Barfield, A. (2005). Complications with collocations: Exploring individual variation in collocation knowledge. *JABAET Journal 9, 85-103.*

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. New York, NY: Springer.

Bauer, L., & Nation, P. (1993). Word families. *International journal of Lexicography, 6*(4), 253-279.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language testing, 27*(1), 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language testing, 16*(2), 131-162.

Beglar, D., & Nation, P. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax—ERRATUM. *Applied Psycholinguistics, 32*(2), 435-452.

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

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research, 74*(4), 245-248.

Bornmann, G., & Munby, I. (2004). Lexical Guessing strategies in EFL reading. Hit or Myth. *Journal of Foreign Language Education, 1,* 3, 23.

Bowne, J. B., Yoshikawa, H., & Snow, C. E. (2017). Relationships of teachers' language and explicit vocabulary instruction to students' vocabulary growth in kindergarten. *Reading Research Quarterly, 52*(1), 7-29.

Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research, 15*(1), 83-97.

Brown, D. (2017). Examining the word family through word lists. *Vocabulary Learning and Instruction, 7* (1), 35–50.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge university press.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a foreign language, 20*(2), 136-163.

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 389.

Chiang, H.-H. (2018). English Vocabulary Size as a Predictor of TOEIC Listening and Reading Achievement among EFL Students in Taiwan. *Theory and Practice in Language Studies, 8*(2), 203-212.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental psychology, 38*(6), 967.

Cobb, T. (2016). Compleat Lexical Tutor (Lextutor). *Retrieved from www.lextutor.ca (last accessed February 2018).*

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior, 21*(4), 451-467.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235.

Danilović, J., Savić, J. D., & Dimitrijević, M. (2013). Affix acquisition order in Serbian EFL learners. *Romanian Journal of English Studies, 10*(1), 77-88.

Davies, M., Wang, X., & Liu, G. (2008). The Corpus of Contemporary American English–A Useful Tool for English Teaching and Research. *Computer-assisted foreign language education in China*, 5, 24-31.

De Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in second language acquisition*, 309-329.

De Graaff, R., & Housen, A. (2009). 38 Investigating the Effects and Effectiveness of L2 Instruction. *The handbook of language teaching*, 726.

de Groot, A. M. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning, 56*(3), 463-506.

de Groot, A. M., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1-56.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning, 55*(S1), 1-25.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In Ross, B. A. (Eds), *Psychology of learning and motivation* (Vol. 53, pp. 63-147). Academic Press.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology, 79*(2), 162.

Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational leadership, 48*(7), 71-76.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In Bjork, E. L., Bjork, R. A. (Eds), *Memory* (pp. 317-344). Academic Press.

Dempster, F. N., & Farris, R. (1990). The spacing effect: Research and practice. *Journal of Research & Development in Education, 23*(2), 97-101.

DeVellis, R. F. (2003). *Scale Development: Theory and Applications* (2nd ed., Vol. 26). Thousand Oaks, CA: Sage Publications.

Dictionaries, O. (2010). *Oxford Essential Arabic Dictionary*. Oxford University Press.

Dictionaries, O., & Attia, M. (2014). *Oxford Arabic Dictionary*. Oxford: Oxford University Press.

Diller, K. C. (1978). *The language teaching controversy*. Rowley, Massachusetts: Newbury House.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*(5), 795.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford: Oxford University Press.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. New York: Routledge.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*(4), 545-565.

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

El-Hibir, B. I., & Al-Taha, F. M. (1992). Orthographic errors of Saudi students learning English. *Language Learning Journal, 5*(1), 85-87.

Ellis, N., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology, 46*(3), 533-558.

Ellis, N. C. (1994). Implicit and explicit processes in language acquisition: An introduction. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp.1-32). San Diego, CA: Academic Press.

Ellis, R. (2001). *Form-focused instruction and second language learning: Language learning monograph*. Wiley-Blackwell.

Eyckmans, J. (2004). *Measuring receptive vocabulary size. Reliability and validity of the Yes/No vocabulary test for French-speaking learners of Dutch* (doctoral dissertation). Retrieved from https://repository.ubn.ru.nl/bitstream/handle/2066/19469/19469_measrevos.pdf

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal, 31*(2), 105-119.

File, K. A., & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? *Tesol Quarterly, 44*(2), 222-249.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EuroSLA Yearbook, 6*(1), 121-145.

Freebody, P., & Anderson, R. C. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior, 15*(3), 19-39.

Fujimori, J. (2005). The lexical composition of two oral communication I textbooks. *The Language Teacher, 29*(7), 15-19.

Gabriel G. (1990) Why a Proper Name has a Meaning: Marty and Landgrebe vs. Kripke. In Mulligan K. (eds) *Mind, Meaning and Metaphysics*. Primary Sources in Phenomenology, vol 3. Springer, Dordrecht

Gairns, R., & Redman, S. (1986). *Working with words: A guide to teaching and learning vocabulary*. Cambridge University Press.

Garrett, P., Griffiths, Y., James, C., & Scholfield, P. (1994). Use of the mother-tongue in second language classrooms: An experimental investigation of effects on the attitudes and writing performance of bilingual UK schoolchildren. *Journal of Multilingual & Multicultural Development*, *15*(5), 371-383.

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in second language acquisition, 21*(2), 319-333.

Giridharan, B. (2010). *An investigative study of English vocabulary acquisition patterns in adult L2 tertiary learners with Chinese/Malay L1* (Doctoral dissertation). Retrieved from https://espace.curtin.edu.au/bitstream/handle/20.500.11937/2416/159516_Giridharan2011.pdf

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*(2), 95-112.

Godden, D. R., & Baddeley, A. D. (1975). Context‐dependent memory in two natural environments: On land and underwater. *British Journal of psychology, 66*(3), 325-331.

González-Fernández, B., & Schmitt, N. (2017). Vocabulary acquisition. *The Routledge handbook of instructed second language acquisition,* 280-298.

Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology, 24*(8), 965-971.

Gor, K., & Vatz, K. (2009). 14 Less Commonly Taught Languages: Issues in Learning and Teaching. *The handbook of language teaching*, 234.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied linguistics, 11*(4), 341-363.

Groot, P. J. (2000). Computer assisted second language vocabulary acquisition. *Language learning & technology, 4*(1), 56-76.

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

Gyllstad, H. (2012, May). Validating the Vocabulary Size Test. A classical test theory approach. In *Poster presented at The Ninth Annual Conference of EALTA, Innsbruck, Austria* (Vol. 31).

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics, 166*(2), 278-306.

Hall, G., & Cook, G. (2012). Own-language use in language teaching and learning. *Language teaching, 45*(3), 271-308.

HAQ, F. & Smadi, O. (1996). Spread of English and westernization in Saudi Arabia. *World Englishes, 15*(3), 307-317.

Harris, V., & Snow, D. (2004). *Doing it for themselves: focus on learning strategies and vocabulary building*. Centre for Information on Language Teaching and Research.

Harwood, N. (2002). Taking a lexical approach to teaching: Principles and problems. *International Journal of Applied Linguistics, 12*(2), 139-155.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House Publishers.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in second language acquisition, 21*(2), 303-317.

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.

Hintzman, D. L. (1976). Repetition and memory. In *Psychology of learning and motivation* (Vol. 10, pp. 47-91). Elsevier.

Hirsch, E. D. (2003). Reading comprehension requires knowledge—of words and the world. *American Educator, 27*(1), 10-13.

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

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review, 56*(2), 308-328.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language testing, 19*(3), 227-245.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. D. M. H. Long (Ed.), *The handbook of second language research* (pp. 349–381). London, England: Blackwell.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in second language acquisition*, *27*(2), 129-140.

Hulstijn, J. H. (2012). Incidental Learning in Second Language Acquisition. In Chapelle, C. A. (Ed.), *The Encyclopaedia of Applied Linguistics*. doi:10.1002/9781405198431.wbeal0136.pub2

Hummel, K. M. (2010). Translation and short-term L2 vocabulary retention: Hindrance or help? *Language Teaching Research, 14*(1), 61-74.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a foreign language, 17*(1), 23-59.

James, C. (1994). Dont shoot my dodo: on the resilience of contrastive and error analysis. *IRAL-International Review of Applied Linguistics in Language Teaching, 32*(3), 179-200.

Javid, C. Z., Al-Asmari, A. R., & Farooq, U. (2012). Saudi undergraduates' motivational orientations towards English language learning along gender and university major lines: A comparative study. *European Journal of Social Sciences, 27*(2), 283-300.

Javid, C. Z., Al-thubaiti, T. S., & Uthman, A. (2013). Effects of English Language Proficiency on the Choice of Language Learning Strategies by Saudi English-Major Undergraduates. *English language teaching, 6*(1), 35-47.

Jeffries, L., Mikulecky, B. S., & Mikulecky, B. (2009). *Basic reading power 1: extensive reading, vocabulary building, comprehension skills, thinking skills*. Pearson Longman.

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in second language acquisition, 24*(4), 617-637.

Jiang, N. (2004). Semantic transfer and its implications for vocabulary teaching in a second language. *The modern language journal, 88*(3), 416-432.

Joe, A. (2010). The Quality and Frequency of Encounters with Vocabulary in an English for Academic Purposes Programme. *Reading in a foreign language*, *22*(1), 117-138.

Jordens, P., & Kellerman, E. (1981). Investigations into the "transfer strategy" in second language learning. In J.-G. Savard and L. Laforge (Eds.), Proceedings of the 5th AILA Congress (pp. 195–215). Laval, Québec: Les Presses de l'Université de Laval.

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction, 36*, 38-45.

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

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365-386.

Keijzer, M. (2007). Last in first out?: An investigation of the regression hypothesis in Dutch emigrants in Anglophone Canada. *Toegepaste Taalwetenschap in Artikelen, 78*(1), 131-139.

Kellerman, E. (1983). Now you see it, now you don't. *Language transfer in language learning, 54*(12), 112-134.

Kellerman, E. (1995). Crosslinguistic influence: Transfer to nowhere? *Annual review of applied linguistics, 15*, 125-150.

Khezrlou, S., Ellis, R., & Sadeghi, K. (2017). Effects of computer-assisted glosses on EFL learners' vocabulary acquisition and reading comprehension in three learning conditions. *System, 65*, 104-116.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The modern language journal, 78*(3), 285-299.

Koda, K. (1997). *Orthographic knowledge in L2 lexical processing A cross-linguistic perspective*. Second Language Vocabulary Acquisition: A Rationale for Pedagogy, 35.

Koirala, C. (2015). The word frequency effect on second language vocabulary learning. Paper presented at *the Critical CALL–Proceedings of the 2015 EUROCALL Conference*, Padova, Italy.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(9), 1297-1317.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*(2), 125-136.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and aging, 25*(2), 498.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly, 13*(4), 377-392.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review, 76*(1), 82.

Laufer, B. (1990). Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals, 23*(2), 147-155.

Laufer B. (1992) How Much Lexis is Necessary for Reading Comprehension? In: Arnaud P.J.L., Béjoint H. (eds) *Vocabulary and Applied Linguistics*. Palgrave Macmillan, London.

Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. *Vocabulary: Description, acquisition and pedagogy*, 140-155.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied linguistics, 19*(2), 255-271.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EuroSLA Yearbook, 5*(1), 223-250.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *Canadian Modern Language Review, 63*(1), 149-166.

Laufer, B. (2017). From word parts to full texts: Searching for effective methods of vocabulary learning. *Language Teaching Research, 21*(1), 5–11

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language testing, 21*(2), 202-226.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied linguistics, 29*(4), 694-716.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied linguistics*, *22*(1), 1-26.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing, 16*(1), 33-51.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), 365-391.

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research, 15*(4), 391-411.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*(1), 89-108.

Lehmann, M. (2007). Is intentional or incidental vocabulary learning more effective. *The International Journal of Foreign Language Teaching, 3*(1), 23-28.

Leminen, A., Lehtonen, M., Bozic, M., & Clahsen, H. (2016). Morphologically complex words in the mind/brain. *Frontiers in human neuroscience, 10*, 47.

Leow, R. P., & Zamora, C. C. (2017). Intentional and incidental L2 learning. *The Routledge handbook of instructed second language acquisition*, 33-49.

Liach, M., & Gallego, M. (2009). Examining the Relationship between Receptive Vocabulary Size and Written Skills of Primary School Learners.. *Journal of the Spanish Association of Anglo-American Studies, 31*(1), 129-147.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science, 25*(3), 639-647.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15*(3), 257-280.

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: the role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning, 7*(3), 175-195.

Lohnas, L. J., & Kahana, M. J. (2014). Compound cuing in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 12.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. *Foreign language research in cross-cultural perspective, 2*(1), 39-52.

Mackey, A., & Gass, S. M. (2013). *Second Language Research: Methodology and Design*. New York: Routledge.

Maguire, M. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). A Unified Theory of Word Learning: Putting Verb Acquisition in Context. *Action meets word: How children learn verbs*, 364.

McKinley, J. I. (2016). Overcoming problematic positionality and researcher objectivity. In McKinley, J. & Rose, H., *Doing Research in Applied Linguistics* (pp. 55-64). New York: Routledge.

McMullen, M. G. (2009). Using language learning strategies to improve the writing skills of Saudi EFL students: Will it really work? *System, 37*(3), 418-433.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods, 45*(2), 499-515.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language testing, 18*(4), 333-349.

Meara, P. (1983). Word associations in a foreign language. *Nottingham Linguistics Circular, 11*(2), 29-38.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. S. a. M. McCarthy (Ed.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 109-121). Cambridge: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language testing, 4*(2), 142-154.

Meara, P. M., & Milton, J. (2003). X-Lex: the Swansea levels test. *Express Publishing*.

Meara, P. (2010). EFL vocabulary tests. *Swansea:_lognostics*. (Original work published in 1992).

Mehrpour, S., & Rahimi, M. (2010). The impact of general and specific vocabulary knowledge on reading and listening comprehension: A case of Iranian EFL learners. *System, 38*(2), 292-300.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. *Vocabulary: Description, acquisition and pedagogy, 33*(2), 84-102.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9*(5), 596-606.

Miles, S., & Kwon, C.-J. (2008). Benefits of using CALL vocabulary programs to provide systematic word recycling. *English Teaching, 63*(1), 199-216.

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412-428.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American, 257*(3), 94-99.

Milton, J. (2006). X-Lex: The Swansea Vocabulary Levels Test. In C. Coombe, P. Davidson and D. Lloyd (eds) *Proceedings of the 7th and 8th Current Trends in English Language testing (CTELT) Conference*, Vol. 4 (pp. 29-39). UAE: TESOL Arabia.

Milton, J. (2008). Vocabulary uptake from informal learning tasks. *Language Learning Journal, 36*(2), 227-237.

Milton, J. (2009). *Measuring second language vocabulary acquisition* (Vol. 45). Multilingual Matters.

Milton, J. (2013). *Measuring the contribution of vocabulary knowledge to proficiency in the four skills*. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.) L2 vocabulary acquisition, knowledge and use (pp. 57-78). Eurosla.

Milton, J., & Hopkins, N. (2005). *Aural Lex. Swansea: Swansea University*.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review, 63*(1), 127-147.

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151-172.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. *Insights into non-native vocabulary teaching and learning*, 83-98.

Mochida, k., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language testing, 23*(1), 73-98.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28*(2), 291-304.

Mondria, J.-A., & Boer, M. W.-D. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign Language. *Applied linguistics, 12*(3), 249-267.

Nagy, W., García, G., Durgunoglu, A., & Hancin-Bhatt, B. (1992). Cross-language transfer of lexical knowledge: bilingual students' use of cognates. *Center for the Study of Reading Technical Report; no. 558*.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American educational research journal, 24*(2), 237-270.

Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. *The nature of vocabulary acquisition, 19*, 35.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). *Learning word meanings from context: How broadly generalizable?* Retrieved from https://www.ideals.illinois.edu/handle/2142/31295

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 233-253.

Nagy, W. E., & Scott, J. A. (2000) Vocabulary processes. *Handbook of reading research, 3*(269-284).

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL, 20*(1), 3-20.

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

Nakata, T., & Suzuki, Y. (2018). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in second language acquisition*, 1-25.

Nakata, T., & Webb, S. (2016). *Does Studying Vocabulary in Smaller Sets Increase Learning*? The Effects of Part and Whole Learning on Second Language Vocabulary Acquisition. *Studies in second language acquisition, 38*(3), 523-552.

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *Tesol Quarterly, 37*(4), 645-670.

Nation, I. (1993). Vocabulary size, growth, and use. *The bilingual lexicon*, 115-134.

Nation, I. (2004). A study of the most frequent word families in the British National Corpus. *Vocabulary in a second language: Selection, acquisition, and testing, 10*, 3-13.

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

Nation, I. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, Milton, J., & Treffers-Daller, J. (Ed.), *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nation, I. S. (2001). *Learning vocabulary in another language*: Ernst Klett Sprachen.

Nation, I. S. (2013). *Teaching & learning vocabulary*: Boston: Heinle Cengage Learning.

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

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, P., & Chung, T. (2009). Teaching and testing vocabulary. In Long M. H. and Doughty C. J. (Eds) *The handbook of language teaching*. John Wiley & Sons

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy, 14*, 6-19.

Newton, J. (1995). Task-based interaction and incidental vocabulary learning: A case study. *Second Language Research, 11*(2), 159-176.

O'Dell, F. (1997). Incorporating vocabulary into the syllabus. *Vocabulary: Description, acquisition and pedagogy*, 258-278.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Oxford, R., & Crookall, D. (1990). Vocabulary learning: A critical analysis of techniques. *TESL Canada Journal*, 09-30.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic bulletin & review, 14*(2), 187-193.

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research, 21*(3), 381-402.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental Vocabulary Acquisition from an Authentic Novel: Do" Things Fall Apart"? *Reading in a foreign language, 22*(1), 31-55.

Peters, E. L2 vocabulary acquisition and reading comprehension: The influence of task complexity. *Investigating tasks in formal language learning*, 178-198.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research, 18*(1), 75-94.

Piasecka, L. (2006). 'Don't lose your head' or how Polish learners of English cope with L2 idiomatic expressions. *Cross-linguistic influences in the second language lexicon*, 246-269.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a foreign language, 18*(1), 1-28.

Pignot-Shahov, V. (2012). Measuring L2 receptive and productive vocabulary knowledge. *Language Studies Working Papers, 4*(1), 37-45.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*(1), 1-19.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70.

Read, J. (2000). *Assessing vocabulary*: Cambridge university press Cambridge.

Read, J. (2004). 7. Research in Teaching Vocabulary. *Annual review of applied linguistics, 24*, 146-161.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*(2), 105-126.

Reynolds, B. L., & Wible, D. (2014). Frequency in incidental vocabulary acquisition research: An undefined concept and some consequences. *Tesol Quarterly, 48*(4), 843-861.

Richards, B., & Malvern, D. (2007). Validity and threats to the validity of vocabulary measurement. *Modelling and assessing vocabulary knowledge*, 79-92.

Richards, B., Malvern, D., & Graham, S. (2008). Word frequency and trends in the development of French vocabulary in lower-intermediate students during Year 12 in English schools. *Language Learning Journal, 36*(2), 199-213.

Richards, J. C. (1976). The role of vocabulary teaching. *Tesol Quarterly*, 77-89.

Rieder, A. (2003). Implicit and explicit learning in incidental vocabulary acquisition. *Views, 12*(2), 24-39.

Ringbom, H. (1978). The influence of the mother tongue on the translation of lexical items. *Interlanguage Studies Bulletin*, 80-101.

Ringbom, H. (1987). *The role of the first language in foreign language learning* (Vol. 34). Philadelphia: Multilingual Matters Ltd.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning* (Vol. 21): Philadelphia: Multilingual Matters Ltd.

Ringbom, H., & Jarvis, S. (2009). The importance of cross-linguistic similarity in foreign language learning. *The handbook of language teaching*, 106-118.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences, 15*(1), 20-27.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Rohrer, D. (2009). Avoidance of overlearning characterises the spacing effect. *European Journal of Cognitive Psychology, 21*(7), 1001-1012.

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

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

Rott, S. (1999). The Effect of Exposure Frequency on Intermediate Language Learners' incidental Vocabulary Acquisition and Retention Through Reading. *Studies in second language acquisition, 21*(4), 589-619.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1161.

Russo, R., Mammarella, N., & Avons, S. (2002). Toward a unified account of spacing effects in explicit cued-memory tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 819.

Saragi, T., Nation, I S.P., Meister, G.F.,. (1978). Vocabulary Learning and Reading. *System, 6*(2), 72-78.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied linguistics, 11*(2), 129-158.

Schmidt, R. (1992). Awareness and second language acquisition. *Annual review of applied linguistics, 13*, 206-226.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. *Implicit and explicit learning of languages, 22*, 165-209.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281-317.

Schmitt, N. (2000). *Vocabulary in language teaching*. Ernst Klett Sprachen.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363.

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

Schmitt, N., & Marsden, R. (2006). *Why is English like that? Historical answers to hard ELT questions*. University of Michigan Press.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge university press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing, 18*(1), 55-88.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *Tesol Quarterly, 36*(2), 145-171.

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

Scholfield, P. (1991). *Vocabulary rate in coursebooks: Living with an unstable lexical economy*. Paper presented at the Proceedings of 5th Symposium on the Description and/or Comparison of English and Greek.

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research, 19*(1), 28-42.

Schwanenflugel, P. J., Akin, C., & Luh, W.-M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20*(1), 96-104.

Schwanenflugel, P. J., Stahl, S. A., & Mcfalls, E. L. (1997). Partial word knowledge and vocabulary growth during reading comprehension. *Journal of Literacy Research, 29*(4), 531-553.

Seabrook, R., Brown, G. D., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied cognitive psychology, 19*(1), 107-122.

Seal, B. D. (1991). Vocabulary learning and teaching. *Teaching English as a second or foreign language, 2,* 296-311.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System, 35*(3), 305-321.

Sheridan, R., & Markslag, L. (2017). Effective Strategies for Teaching Vocabulary: An Introduction to Engaging Cooperative Vocabulary Card Activities. *PASAA: Journal of Language Teaching and Learning in Thailand, 53,* 214-229.

Shillaw, J. (1996). The application of Rasch modelling to yes/no vocabulary tests. *Vocabulary Acquisition Research Group*. Retrieved from http://www.lognostics.co.uk/vlibrary/shillaw1996.doc

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 907.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied cognitive psychology, 25*(5), 763-767.

Sonbul, S., & Schmitt, N. (2009). Direct teaching of vocabulary after reading: Is it worth the effort? *ELT journal, 64*(3), 253-260.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in second language acquisition, 31*(4), 577-607.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139-152.

Stockwell, R. P., Bowen, J. D., & Martin, J. W. (1965). *The grammatical structures of English and Spanish* (Vol. 4). University of Chicago Press.

Summers, D., & Stock, P. (1992). *Longman dictionary of English language and culture*. Longman.

Tang, E., & Nesi, H. (2003). Teaching vocabulary in two Chinese classrooms: Schoolchildren’s exposure to English words in Hong Kong and Guangzhou. *Language Teaching Research, 7*(1), 65-97.

Tekmen, E. A. F., & Daloğlu, A. (2006). An investigation of incidental vocabulary acquisition in relation to learner proficiency level and word frequency. *Foreign Language Annals, 39*(2), 220-243.

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15*(5), 529-536.

Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning, 59*(3), 623-646.

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 28*(3), 437.

Wallace, M. J. (2006). *Action research for language teachers*. Ernst Klett Sprachen.

Walsh, M., Gluck, K., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J., Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General, 147*(9), 1325.

Wang, X., Davies, M., & LIU, G. (2008). A Good Platform for English Teachers and Learners: The Corpus of Contemporary American English (COCA). *Computer-Assisted Foreign Language Education in China, 5*.

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata* (1), 53-68.

Waring, R., & Takaki, (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a foreign language, 15*(2), 130.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied linguistics, 28*(1), 46-65.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in second language acquisition, 30*(1), 79-95.

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal, 44*(3), 263-277.

West, M. P. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*: Longmans, Green.

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal, 40*(1), 125-137.

Wozniak, R. H. (1999). Introduction to memory: Hermann Ebbinghaus (1885/1913). *Classics in the history of psychology*.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System, 35*(2), 182-191.

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.

Appendices

Appendix A: **Frequency profiles of vocabulary in the exercises using Lextutor BNC-20**

level	**tokens**	**cumul%**
K-1	86.50	86.50
K-2	8.39	94.89
K-3	4.01	98.90
K-4	0.36	99.26
K-5	0.73	99.99
OFF	0.00	≈100%

Exercise 1A

level	**tokens**	**cumul%**
K-1	88.13	88.13
K-2	7.19	95.32
K-3	3.60	98.92
K-4	0.36	99.28
K-5	0.72	100.00
OFF	0.00	≈100%

Exercise 1B

level	**tokens**	**cumul%**
K-1	84.13	84.13
K-2	7.69	91.82
K-3	5.77	97.59
K-4	1.44	99.03
K-5	0.96	99.99
OFF	0.00	≈100%

Exercise 2A

level	**tokens**	**cumul%**
K-1	83.96	83.96
K-2	9.63	93.59
K-3	4.28	97.87
K-4	1.60	99.47
K-5	0.53	100.00
OFF	0.00	≈100%

Exercise 2B

Appendix B: **learners' perception questionnaire (English)**

Name: .... **ID:** ....

Please tell us your opinion about the following statements using the following five-point scale, circle your best choice clearly after each statement.

Q1: What do you think of the following statements based on your vocabulary learning in comparison between spaced practice and massed practice?

(1) Strongly disagree	(2) Disagree	(3) Neutral	(4) Agree	(5) Strongly agree

A. I memorize words better in spaced learning than massed learning. 1 2 3 4 5
B. I memorize more words in spaced learning than massed learning. 1 2 3 4 5
C. I retain words better in spaced learning than massed learning. 1 2 3 4 5
D. I recall words better in spaced learning than massed learning. 1 2 3 4 5
E. I learn words quickly in spaced learning than massed learning. 1 2 3 4 5
F. I can focus better in spaced learning than massed learning. 1 2 3 4 5
G. I feel motivated in spaced learning than massed learning. 1 2 3 4 5
H. I feel less bored in spaced learning than massed learning

1 2 3 4 5
I. I prefer spaced learning than massed in learning new words. 1 2 3 4 5
J. What is your thoughts on learning vocabulary using spaced practice or massed practice? Please say …………………………………………………………………………………………………… …………………………………………………………………………………………………… ……………………………………………………………………………………………………

Appendix C: **learners' perception questionnaire (Arabic)**

استبيان

الاسم: .... **الرقم الجامعي**: ....

عزيزي الطالب نرجو منك أن تخبرنا عن رأيك في العبارات التالية وذلك بوضع دائرة على الرقم الذي يتفق مع رأيك أمام كل عبارة.

س: ما رأيك في العبارات التالية بناء على تعلمك للمفردات من خلال المقارنة بين التعلم المتباعد والتعلم المتتابع؟

	(1) **لا أوافق بشدة**	**(2)** **لا أوافق**	**(3)** **محايد**	**(4)** **أوافق**	**(5)** **أوافق بشدة**
أحفظ الكلمات بشكل أفضل عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
أحفظ كلمات أكثر عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
يرسخ حفظ الكمات بشكل أفضل عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
أتذكر الكلمات بشكل أفضل في الاختبارات عند تعلمها عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
تتطور مفرداتي اللغوية بشكل أفضل عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
أتعلم بسرعة عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5
أستطيع التركيز بشكل أفضل عندما يكون التعلم عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	1	2	3	4	5

اشعر بتحفيز أكبر عندما يكون تعلم المفردات عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	5 4 3 2 1
أشعر بالملل عندما يكون تعلم المفردات عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	5 4 3 2 1
أفضل تعلم الكلمات الجديدة عن طريق التكرار المتباعد مقارنة بالتكرار المتتابع	5 4 3 2 1
الرجاء التكرم بكتابة آراء أخرى لتعلمك المفردات الإنجليزية باستخدام التكرار المتباعد والمتكرر لم تشر اليها في الطرق المذكورة أعلاه.	

Appendix D: **Biodata and language learning history questionnaire (English)**

Name: ....

Age: ....

Contact number: ....

Email: ....

Please answer the following questions (you are not obliged to answer any question).

How many years have you been learning English?

....

Have you studied English abroad? *If yes, where and how long?*

....

How many lectures do you have in your schedule today? <u>Please mention name of course and time</u>.

....

....

How many hours do you study this semester?

....

Do you have part-time or fulltime job while studying? *If yes, when and how long?*

....

....

Do you have daily commitments or activities which affect your studies? *If yes, what, when and how long?*

....

....

Do you suffer from any medical problems that affect your studies? *If yes, please explain.*

....

....

Do you suffer or previously suffered from memory loss or memory issues? *If yes, please explain.*

....

....

Thank you for your participation!

Appendix E: **Biodata and language learning history questionnaire (Arabic)**

أسئلة استطلاعية

الاسم: ....

العمر: ....

رقم الجوال: ....

عنوان البريد الإلكتروني: ....

الرجاء الإجابة على الأسئلة التالية (أنت لست مضطرا للإجابة على أي سؤال).

كم عدد سنوات تعلمك للغة الإنجليزية؟

....

هل درست اللغة الإنجليزية في الخارج؟ إذا كانت الإجابة بنعم، أين ومتى؟

....

كم عدد المحاضرات التي لديك اليوم؟ يرجى ذكر اسم المقرر ووقت المحاضرة.

....

كم عدد الساعات الدراسية لديك لهذا الفصل؟

....

هل لديك وظيفة بدوام جزئي أو كامل أثناء الدراسة؟ إذا كان الجواب نعم، وضح طبيعة العمل ووقته؟

....

....

هل لديك التزامات أو أنشطة إضافية تؤثر على دراستك الجامعية؟ إذا كانت الإجابة بنعم، ماهي ومتى وما مدتها؟

....

....

هل تعاني من أي مشاكل طبية تؤثر على دراستك؟ إذا كانت الإجابة بنعم، رجاء التوضيح.

....

....

هل تعاني أو عانيت مسبقا من مشاكل في الذاكرة أو فقدان الذاكرة؟ إذا كانت الإجابة بنعم، رجاء التوضيح

....

....

Appendix F: **Translation Multiple-choice pre-test**

Name: ………………………………..…………….……… University ID: ………………… Class/Division ………

Please choose the correct Arabic translation of each English word. If you know a translation that is not provided, please tick box (e).
اختر الترجمة العربية الصحيحة لكل الكلمات الإنجليزية التالية، وفي حالة معرفتك بترجمة غير المزودة في الخيارات ضع علامة Xعند مربع الخيار (e).

1. thread

a. إشارة	b. غيمة	c. خيط	d. شعار	e.

2. breeze

a. عصير	b. ستارة	c. غروب	d. نسيم	e.

3. palace

a. قصر	b. ممحاة	c. مقلمة	d. عبارة	e.

4. defeat

a. عمارة	b. وليمة	c. هزيمة	d. سفارة	e.

5. sleeve

a. مسطح	b. كم	c. دعامة	d. سفير	e.

6. helmet

a. خوذة	b. مصنع	c. مؤشر	d. قارب	e.

7. collar

a. خليوي	b. منضدة	c. معصم	d. ياقة	e.

8. debris

a. سعال	b. حطام	c. سجاد	d. مرسام	e.

9. permit

a. لوحة	b. منتدى	c. تصريح	d. مأتم	e.

10. trauma

a. صدمة	b. منفاخ	c. منفى	d. مسرح	e.

11. escape

a. هجاء	b. سمر	c. ثراء	d. هروب	e.

12. vendor

a. ملعب	b. بائع	c. صالة	d. مصنع	e.

13. mentor

a. حلوى	b. نعناع	c. مرشد	d. مكسب	e.

14. refuge

a. دعاية	b. وقود	c. ملجأ	d. شاعر	e.

15. intent

a. يتيم	b. دولاب	c. نية	d. مدى	e.

16. glimpse

a. لمحة	b. غمزة	c. مساحة	d. ملتقى	e.

17. hostage

a. حمولة	b. مأهول	c. سرية	d. رهينة	e.

18. vaccine

a. متوفر	b. لقاح	c. متجر	d. سفير	e.

19. consent

a. مسكن	b. رحيق	c. موافقة	d. عاصفة	e.

20. fatigue

a. سنام	b. إرهاق	c. شريحة	d. دسم	e.

21. verdict

a. مرهم	b. حكم	c. سهر	d. قاحل	e.

22. texture

a. نص	b. مرآب	c. سجن	d. نسيج	e.

23. terrain

a. قنطار	b. تربة	c. غمرة	d. تضاريس	e.

24. sponsor

a. كفيل	b. رفيق	c. جاسوس	d. مسؤول	e.

25. nominee

a. إطار	b. مرشح	c. شفاف	d. ملجأ	e.

26. dignity

a. كرامة	b. بطانة	c. رماية	d. سلامة	e.

27. dilemma

a. وهم	b. طاغية	c. معضلة	d. مساهمة	e.

28. pioneer

a. سجين	b. مسموم	c. عجين	d. ريادي	e.

29. density

a. كثافة	b. أشعة	c. قناع	d. سد	e.

30. harmony

a. موسيقى	b. ألفة	c. شعلة	d. منطقة	e.

31. plunge

a. يغطس	b. يدمر	c. يتسلى	d. يناجي	e.

32. thrive

a. يسرق	b. يزدهر	c. يتمنى	d. يساهم	e.

33. shrink

a. يسقي	b. يعصر	c. ينكمش	d. يتدلى	e.

34. object

a. يحدد	b. يوضح	c. يلوح	d. يعترض	e.

35. donate

a. يسقي	b. يساير	c. يتبرع	d. يتناقش	e.

36. insert

a. يصيد	b. يدخل	c. يبيد	d. ينمو	e.

37. foster

a. يستثمر	b. يشترك	c. يتمرن	d. يربي	e.

38. compel

a. يصوم	b. يعين	c. يتلون	d. يرغم	e.

39. battle

a. يتعارك	b. يتشقق	c. يدور	d. يجدف	e.

40. invade

a. يدعو	b. يجتاح	c. يحتفل	d. يخطط	e.

41. comply

a. يطيع	b. يتقدم	c. يهمش	d. يرعى	e.

42. mutter

a. يقسم	b. يحدد	c. يتمتم	d. ينحت	e.

43. forbid

a. يلعب	b. يراهن	c. ينهى	d. ينهش	e.

44. reward

a. يعود	b. يكافئ	c. يتفكر	d. يتعبد	e.

45. regret

a. يندم	b. يهنأ	c. يخطط	d. يتراجع	e.

46. stumble

a. يضرب	b. ينثر	c. يلحن	d. يتعثر	e.

47. condemn

a. يتهم	b. يشجب	c. يلتهم	d. يواسي	e.

48. persist

a. يسبق	b. يصر	c. يستعد	d. يبدد	e.

49. exploit

a. يطير	b. ينير	c. يبحث	d. يستغل	e.

50. disturb

a. ينثر	b. يزعج	c. يحفر	d. يلقح	e.

51. exhaust

a. يتعب	b. يعتذر	c. يتجبر	d. يغلف	e.

52. provoke

a. يثبت	b. ينسخ	c. يستخير	d. يستفز	e.

53. dictate

a. يسيطر	b. يلقن	c. يتأمر	d. يغوص	e.

54. descend

a. يرسل	b. ينحدر	c. يهاجر	d. يحتقر	e.

55. concede

a. يقر	b. يضاعف	c. يطرح	d. يعارض	e.

56. prevail

a. يغطي	b. يتكيف	c. يجرح	d. يغلب	e.

57. harvest

a. يبارز	b. يحصد	c. يناور	d. يسيطر	e.

58. inherit

a. يحرث	b. يناضل	c. يرث	d. يسكب	e.

59. isolate

a. يستعد	b. يحتضر	c. يعزل	d. يجامل	e.

60. utilize

a. يبرمج	b. يسير	c. يسخر	d. يطور	e.

Thank you for your participation!

Appendix G: **Vocabulary Achievement Test of Strength (List A)**

Name: ....

UNID: ....

PART ONE *active recall*

Please supply the English translation for each Arabic word below.

1. يتعثر ....
2. رهينة ....
3. يدخل ....
4. يعترض ....
5. نسيم ....
6. ياقة ....
7. يلقن ....
8. كم ....
9. موافقة ....
10. يزدهر ....
11. يغطس ....
12. يصر ....
13. يشجب ....
14. لقاح ....
15. إرهاق ....
16. خوذة ....
17. يتبرع ....
18. قصر ....
19. يربي ....
20. يستفز ....
21. ينكمش ....
22. يستغل ....
23. تضاريس ....
24. يزعج ....
25. هزيمة ....
26. حكم ....
27. نسيج ....
28. لمحة ....
29. خيط ....
30. يتعب ....

Name: ……………………………………………………………………………………………

UNID: ………………………

PART TWO *passive recall*

Please translate the English words on the left into Arabic.

Name: ……………………………………………………………………………………………

1. **consent** ……………......
2. **defeat** ……………......
3. **condemn** ……………......
4. **collar** ……………......
5. **stumble** ……………......
6. **vaccine** ……………......
7. **object** ……………......
8. **shrink** ……………......
9. **verdict** ……………......
10. **insert** ……………......
11. **provoke** ……………......
12. **exhaust** ……………......
13. **fatigue** ……………......
14. **sleeve** ……………......
15. **helmet** ……………......
16. **plunge** ……………......
17. **thread** ……………......
18. **terrain** ……………......
19. **persist** ……………......
20. **thrive** ……………......
21. **breeze** ……………......
22. **palace** ……………......
23. **disturb** ……………......
24. **exploit** ……………......
25. **hostage** ……………......
26. **foster** ……………......
27. **donate** ……………......
28. **texture** ……………......
29. **dictate** ……………......
30. **glimpse** ……………......

UNID: ………………………

PART THREE *active recognition*

Please choose the English translation for each Arabic word. You may choose only one option.

Example:
طعام
[✓ a. food] b. water c. football d. street

1. أصر
a. sustain b. persist c. satisfy d. consume

2. استزف
a. educate b. qualify c. exhaust d. impress

3. حكم
a. verdict b. custom c. scheme d. galaxy

4. خوذة
a. barrel b. stroke c. valley d. helmet

5. انكمش
a. shrink b. explode c. request d. undergo

6. ياقة
a. legacy b. collar c. ethics d. summit

7. لقاح
a. powder b. cookie c. rhythm d. vaccine

8. ملمس
a. script b. tactic c. texture d. margin

9. خيط
a. thread b. square c. tunnel d. virtue

10. نسيم
a. ritual b. insect c. breeze d. combat

11. هزيمة
a. suburb b. defeat c. motive d. temple

12. يعزز
a. appoint b. correct c. portray d. foster

13. قصر
a. domain b. strain c. colony d. palace

14. استفز
a. exhibit b. recruit c. provoke d. compose

15. موافقة
a. consent b. glance c. ballot d. excuse

16. لقن
a. concern b. dictate c. finance d. suppose

17. لمحة

a. canvas b. glimpse c. format d. spouse

18. تضاريس

a. terrain b. needle c. pastor d. equity

19. إرهاق

a. statue b. parade c. fatigue d. handle

20. استغل

a. entitle b. exclude c. reverse d. exploit

21. ازدهر

a. confuse b. thrive c. swallow d. enforce

22. أزعج

a. disturb b. contend c. protest d. consult

23. غطس

a. forgive b. project c. execute d. plunge

24. ادخل

a. average b. trigger c. insert d. convict

25. تعثر

a. stumble b. scatter c. relieve d. suspend

26. رهينة

a. hostage b. entity c. divorce d. uniform

27. اعترض

a. rebuild b. object c. endorse d. program

28. تبرع

a. picture b. donate c. confess d. differ

29. شجب

a. damage b. derive c. condemn d. prompt

30. كم

a. scandal b. contest c. suspect d. sleeve

Name: ....

UNID: ....

PART FOUR *passive recognition* (List A)

Please choose the Arabic translation for each English word. You may choose only one option.

Example:
food
b. ماء   b. ملح   [✓ c. طعام]   d. لوحة

1. **glimpse**
a. سلسلة   b. وحدة   c. لمحة   d. حالة

2. **condemn**
a. شجب   b. أمد   c. أصاب   d. ظهر

3. **collar**
a. قمة   b. ياقة   c. مقطورة   d. أسلوب

4. **exploit**
a. تربص   b. ركب   c. طبع   d. استغل

5. **sleeve**
a. طبقة   b. كم   c. مدير   d. حبكة

6. **texture**
a. مستوطنة   b. ملمس   c. مركبة   d. إبرة

7. **shrink**
a. انكمش   b. ألقى   c. وجد   d. أطعم

8. **object**
a. صرف   b. شارك   c. منع   d. اعترض

9. **consent**
a. موافقة   b. قفص   c. قلق   d. نسخة

10. **hostage**
a. قطيع   b. متسول   c. رهينة   d. قطعة

11. **insert**
a. فرق   b. أخرج   c. طلق   d. ادخل

12. **persist**
a. صدق   b. أجتنب   c. أصر   d. أمر

13. **defeat**
a. حظيرة   b. وادي   c. أخلاق   d. هزيمة

14. **fatigue**
a. منتدى   b. غابة   c. خادم   d. إرهاق

15. **donate**
a. تبرع   b. أخذ   c. حلف   d. كشف

16. **exhaust**
a. ذكر   b. استنزف   c. صنع   d. نزع

17. **provoke**
a. حلل   b. استفز   c. أشرك   d. أسرع

18. **plunge**

a. عرض b. كذب c. غطس d. يسر

19. **breeze**

a. نسيم b. سراب c. أسير d. عضوي

20. **thrive**

a. دخل b. لبس c. ازدهر d. جحد

21. **helmet**

a. جراح b. تحقيق c. سلة d. خوذة

22. **vaccine**

a. لقاح b. ولي c. جوهر d. صياد

23. **foster**

a. أعلن b. يعزز c. طهر d. رمى

24. **thread**

a. عذر b. جيش c. خيط d. كوخ

25. **stumble**

a. تعثر b. أتبع c. فرض d. هلك

26. **dictate**

a. سأم b. أنطلق c. ثبت d. لقن

27. **verdict**

a. سيل b. منطقة c. حكم d. سوق

28. **palace**

a. سعادة b. قصر c. مزاد d. معرض

29. **disturb**

a. أزعج b. سمع c. فعل d. عدل

30. **terrain**

a. نصيب b. مجمع c. تضاريس d. نكهة

Appendix H: **Vocabulary Achievement Test of Strength (List B)**

Name: ....

UNID: ………………………

PART ONE *active recall*

Please supply one English translation for each Arabic word below.

1. يندم ………………………………

2. كفيل ………………………………

3. ينهى ………………………………

4. يطيع ………………………………

5. تصريح ………………………………

6. ملجأ ………………………………

7. يسخر ………………………………

8. صدمة ………………………………

9. كرامة ………………………………

10. تعارك ………………………………

11. يرغم ………………………………

12. يقر ………………………………

13. ينحدر ………………………………

14. مرشح ………………………………

15. معضلة ………………………………

16. مرشد ………………………………

17. يتمتم ………………………………

18. بائع ………………………………

19. يكافأ ………………………………

20. يعزل ………………………………

21. يجتاح ………………………………

22. يغلب ………………………………

23. ألفة ………………………………

24. يحصد ………………………………

25. هروب ………………………………

26. ريادي ………………………………

27. كثافة ………………………………

28. نية ………………………………

29. حطام ………………………………

30. يرث ………………………………

Name: ………………………………………………………………………………………

UNID: ………………………

PART TWO *passive recall*

Please translate the English words on the left into Arabic.

Name: ....

1. **dignity** ....
2. **escape** ....
3. **descend** ....
4. **refuge** ....
5. **regret** ....
6. **nominee** ....
7. **comply** ....
8. **invade** ....
9. **pioneer** ....
10. **forbid** ....
11. **isolate** ....
12. **inherit** ....
13. **dilemma** ....
14. **vendor** ....
15. **mentor** ....
16. **compel** ....
17. **debris** ....
18. **harmony** ....
19. **concede** ....
20. **battle** ....
21. **permit** ....
22. **trauma** ....
23. **harvest** ....
24. **prevail** ....
25. **sponsor** ....
26. **reward** ....
27. **mutter** ....
28. **density** ....
29. **utilize** ....
30. **intent** ....

UNID: ....

PART THREE *active recognition* (List B)

Please choose the English translation for each Arabic word. You may choose only one option.

Example:
طعام
c. food b. water c. football d. street

1. أقر

a. submit b. confess c. concede d. prompt

2. ورث

a. inherit b. finance c. murder d. trigger

3. ريادي

a. tunnel b. venture c. pioneer d. format

4. مرشد

a. pursuit b. crystal c. ritual d. mentor

5. اجتاح

a. inspire b. invade c. signal d. entitle

6. ملجأ

a. complex b. refuge c. racism d. suspect

7. مرشح

a. nominee b. uniform c. valley d. glance

8. كثافة

a. density b. surgeon c. warrior d. contest

9. حطام

a. divorce b. miracle c. arrival d. debris

10. تصريح

a. stretch b. permit c. motive d. square

11. هروب

a. conduct b. legacy c. escape d. habitat

12. كافأ

a. differ b. impress c. forgive d. reward

13. صدمة

a. trauma b. spouse c. needle d. parade

14. عزل

a. govern b. rescue c. derive d. isolate

15. كرامة

a. exhibit b. pastor c. dignity d. suburb

16. سخر

a. confess b. utilize c. resign d. depict

17. نية

a. rhythm b. shuttle c. primary d. intent

18. الفة

a. domain b. gallery c. harmony d. grocery

19. معضلة

a. barrel b. dilemma c. inquiry d. mandate

20. غلب

a. prevail b. desire c. swallow d. concern

21. تعارك

a. battle b. dismiss c. devote d. portray

22. حصد

a. explode b. convert c. witness d. harvest

23. أرغم

a. convict b. scatter c. compel d. endorse

24. نهى

a. pretend b. forbid c. relieve d. confuse

25. ندم

a. contend b. regret c. modify d. excuse

26. كفيل

a. sponsor b. monster c. courage d. entity

27. أطاع

a. educate b. survey c. justify d. comply

28. تمتم

a. bounce b. endure c. mutter d. picture

29. انحدر

a. squeeze b. debate c. compose d. descend

30. بائع

a. margin b. vendor c. playoff d. pension

Name: ....

UNID: ....

PART FOUR *passive recognition* (List B)

Please choose the Arabic translation for each English word. You may choose only one option.

Example:
food
a. ماء b. ملح c. طعام d. لوحة

1. **intent**
a. نية b. مأساة c. راحة d. رفيق

2. **descend**
a. لبس b. انحدر c. سلك d. ظلم

3. **refuge**
a. حساب b. شراكة c. ملجأ d. إصابة

4. **prevail**
a. أخلف b. قعد c. أنفق d. غلب

5. **vendor**
a. بائع b. مؤسسة c. مجلس d. احتياط

6. **density**
a. صندوق b. محصول c. متحف d. كثافة

7. **invade**
a. بخل b. نظر c. أكل d. اجتاح

8. **comply**
a. صرف b. أطاع c. نبذ d. أرتاب

9. **dignity**
a. صيغة b. كرامة c. رصيف d. كتيبة

10. **sponsor**
a. كفيل b. خطة c. سراب d. أخدود

11. **forbid**
a. قطع b. خاض c. نهى d. غلب

12. **concede**
a. ورث b. أعتدى c. أضاع d. اقر

13. **escape**
a. وزير b. نهاية c. هروب d. مرعى

14. **dilemma**
a. أفق b. عريشة c. معضلة d. سرداب

15. **mutter**
a. تمتم b. أخر c. وقى d. تمتع

16. **inherit**
a. غفر b. ورث c. زعم d. هاجر

17. **isolate**
a. عزل b. وعظ c. أرسل d. وقع

18. **compel**
a. سكن b. بنى c. أرغم d. أفسد

19. **permit**
a. قرار b. تصريح c. بارود d. مطل

20. **battle**
a. آستئذن b. رجع c. تبع d. تعارك

21. **mentor**
a. مرشد b. شرف c. مخرج d. جامع

22. **nominee**
a. مرشح b. عادة c. سور d. فضيلة

23. **reward**
a. كافأ b. خاف c. أيد d. سخر

24. **debris**
a. جاذبية b. مسيرة c. طلاق d. حطام

25. **regret**
a. فسق b. ندم c. أطاع d. أناب

26. **utilize**
a. أقبل b. فطر c. قذف d. سخر

27. **pioneer**
a. شيخ b. ملاحظة c. ريادي d. محارب

28. **trauma**
a. جنين b. صدمة c. ابتلاء d. إيراد

29. **harvest**
a. وعد b. تقبل c. حصد d. أطمأن

30. **harmony**
a. مؤتمر b. الفة c. جماعة d. مبارزة

Appendix I: **Treatment Exercises**

EXERCISE ONE: Multiple-Choice Exercise (List A)

1.	a stretch of land, especially with regard to its physical features			
	a. terrain	b. custom	c. scheme	d. galaxy
2.	make (someone) feel very tired			
	a. sustain	b. exhaust	c. satisfy	d. consume
3.	a substance that is usually injected into a person or animal to protect against a particular disease			
	a. vaccine	b. barrel	c. stroke	d. valley
4.	the fact of losing against someone in a fight or competition			
	a. scandal	b. contest	c. suspect	d. defeat
5.	a large house that is the official home of a king, queen, or other person of high social rank			
	a. legacy	b. ethics	c. palace	d. summit
6.	a person seized or held as security for the fulfilment of a condition			
	a. powder	b. cookie	c. hostage	d. rhythm
7.	trip or momentarily lose one's balance; almost fall			
	a. explode	b. request	c. undergo	d. stumble
8.	to criticize something or someone strongly, usually for moral reasons			
	a. condemn	b. appoint	c. correct	d. portray
9.	a light and pleasant wind			
	a. script	b. breeze	c. tactic	d. margin
10.	to speak or read (something) to a person who writes it down or to a machine that records it			
	a. exhibit	b. dictate	c. recruit	d. compose
11.	the way that something feels when you touch it			
	a. texture	b. square	c. tunnel	d. virtue
12.	the part around the neck of a piece of clothing			
	a. ritual	b. insect	c. collar	d. combat
13.	a strong, hard hat that covers and protects the head			
	a. suburb	b. helmet	c. motive	d. temple
14.	to disagree with something or oppose something			
	a. concern	b. finance	c. suppose	d. object
15.	a long, thin piece of cotton, silk, etc., used for sewing			
	a. domain	b. strain	c. colony	d. thread
16.	to give money or goods to help a person or organization			
	a. donate	b. entitle	c. exclude	d. reverse
17.	permission for something to happen or be done			
	a. glance	b. ballot	c. excuse	d. consent

18.	to interrupt or bother (someone or something)			
	a. contend	b. disturb	c. swallow	d. consult
19.	to fall or jump suddenly from a high place			
	a. confuse	b. protest	c. plunge	d. enforce
20.	to see something or someone for a very short time or only partly			
	a. glimpse	b. entity	c. divorce	d. uniform
21.	to make or try to make a person or an animal angry			
	a. average	b. relieve	c. convict	d. provoke
22.	to grow or develop successfully			
	a. scatter	b. trigger	c. thrive	d. suspend
23.	to become smaller, or to make something smaller			
	a. rebuild	b. endorse	c. shrink	d. program
24.	a judgment or opinion about something			
	a. canvas	b. verdict	c. format	d. spouse
25.	to use something in a way that helps you			
	a. exploit	b. confess	c. differ	d. damage
26.	to continue in an opinion or course of action in spite of difficulty or opposition			
	a. picture	b. derive	c. persist	d. prompt
27.	to help (something) grow or develop			
	a. forgive	b. project	c. execute	d. foster
28.	a part of a garment that covers an arm			
	a. needle	b. sleeve	c. pastor	d. equity
29.	place, fit, or push (something) into something else			
	a. educate	b. qualify	c. insert	d. impress
30.	the state of being very tired			
	a. statue	b. fatigue	c. parade	d. handle

EXERCISE ONE: Multiple-Choice Exercise (List B)

1.	the state or quality of being worthy of honour or respect			
	a. handful	b. horizon	c. dignity	d. edition
2.	cause (a person or place) to be or remain alone or apart from others			
	a. absorb	b. isolate	c. assert	d. wander
3.	to act according to an order, set of rules, or request			
	a. comply	b. complex	c. depict	d. borrow
4.	broken or torn pieces of something larger			
	a. debris	b. submit	c. arrival	d. cluster
5.	a person who has been suggested for an election or job.			
	a. habitat	b. actress	c. carrier	d. nominee
6.	to pick and collect crops, or to collect plants, animals, or fish to eat			
	a. devote	b. admire	c. harvest	d. exceed
7.	a severe emotional shock and pain caused by an extremely upsetting experience			
	a. miracle	b. trauma	c. charity	d. venture
8.	to refuse to allow something			
	a. resign	b. invent	c. murder	d. forbid
9.	an experienced and trusted adviser			
	a. grocery	b. blanket	c. courage	d. mentor
10.	to win over an opponent especially in a long or difficult contest			
	a. excuse	b. render	c. prevail	d. endure
11.	to enter (a place) in large numbers			
	a. invade	b. behave	c. modify	d. survey
12.	a difficult situation or problem.			
	a. formula	b. dilemma	c. gallery	d. inquiry
13.	someone who is selling something			
	a. vendor	b. surgeon	c. counsel	d. essence
14.	an official document that allows you to do something or go somewhere			
	a. servant	b. gravity	c. racism	d. permit

15.	a person taking official responsibility for the actions of another			
	a. pitcher	b. stretch	c. sponsor	d. hallway
16.	the state of being in agreement			
	a. monster	b. harmony	c. crystal	d. warrior
17.	the quantity of people or things in a given area or space.			
	a. curtain	b. density	c. monitor	d. playoff
18.	(a place that gives) protection or shelter from danger, trouble, unhappiness, etc.			
	a. landing	b. refuge	c. conduct	d. vitamin
19.	to feel sad or sorry about (something that you did or did not do)			
	a. regret	b. convey	c. desire	d. resume
20.	to give something in exchange for good behaviour or good work, etc			
	a. bounce	b. signal	c. rescue	d. reward
21.	to receive money, a house, etc. from someone after they have died			
	a. inherit	b. punish	c. cancel	d. debate
22.	to go or come down			
	a. vanish	b. dismiss	c. respect	d. descend
23.	a person who is one of the first people to do something.			
	a. mandate	b. reserve	c. pioneer	d. pension
24.	to engage in a fight or struggle against			
	a. justify	b. possess	c. battle	d. squeeze
25.	to force someone to do something			
	a. convert	b. pretend	c. violate	d. compel
26.	to say something in a low or barely audible voice			
	a. witness	b. inspire	c. mutter	d. assign
27.	the fact that you want and plan to do something			
	a. shuttle	b. intent	c. exhibit	d. liberty
28.	to use something in an effective way			
	a. occupy	b. retain	c. utilize	d. confess
29.	to admit, often unwillingly, that something is true			
	a. concede	b. injure	c. govern	d. praise
30.	the act of successfully getting out of a place or a dangerous or bad situation			
	a. escape	b. garbage	c. primary	d. pursuit

EXERCISE TWO: Fill-in-the-gap (List A)

1.	Some plants ………… in shade, while others will not.			
	a. thrive	b. sustain	c. satisfy	d. consume
2.	if you pull a ………… from the sock, it will ruin.			
	a. custom	b. scheme	c. galaxy	d. thread
3.	Don't let this ………… you any more than it already has.			
	a. educate	b. disturb	c. qualify	d. impress
4.	The royal ………… is open to the public.			
	a. barrel	b. stroke	c. palace	d. valley
5.	The jury returned a ………… of guilty.			
	a. verdict	b. legacy	c. ethics	d. summit
6.	What did you say to ………… him?			
	a. provoke	b. explode	c. request	d. undergo
7.	Would anyone ………… if we started the meeting now?			
	a. appoint	b. correct	c. portray	d. object
8.	They were held ………… by armed rebels.			
	a. contest	b. hostage	c. cookie	d. rhythm
9.	After their ………… in battle, the soldiers surrendered.			
	a. script	b. powder	c. virtue	d. defeat
10.	This journal offers a ………… of his life as a child.			
	a. scandal	b. glimpse	c. tactic	d. margin
11.	Be careful not to ………… on the uneven pavement.			
	a. concern	b. finance	c. stumble	d. suppose
12.	This ………… protects against some kinds of the bacteria.			
	a. square	b. tunnel	c. vaccine	d. suspect
13.	If you ………… with this behaviour, you will be punished.			
	a. entitle	b. exclude	c. reverse	d. persist
14.	………… your credit card here.			
	a. Insert	b. Confuse	c. Swallow	d. Enforce

15.	It's very risky to ride your bike without wearing a …………..			
	a. ritual	b. insect	c. combat	d. helmet
16.	We need to …………. this opportunity.			
	a. contend	b. protest	c. exploit	d. consult
17.	They were unable to continue the race due to …………..			
	a. suburb	b. fatigue	c. temple	d. domain
18.	The …………. rippled the water.			
	a. breeze	b. motive	c. strain	d. colony
19.	He wore a shirt with a tight-fitting …………..			
	a. glance	b. collar	c. ballot	d. excuse
20.	That sweater will …………. if you wash it.			
	a. exhibit	b. shrink	c. recruit	d. compose
21.	…………. the books to a library.			
	a. donate	b. forgive	c. project	d. execute
22.	We …………. all acts of terrorism.			
	a. average	b. convict	c. scatter	d. condemn
23.	Such conditions …………. the spread of the disease.			
	a. confess	b. foster	c. rebuild	d. picture
24.	She wiped her nose on her …………..			
	a. canvas	b. format	c. sleeve	d. needle
25.	They can't publish your name without your …………..			
	a. spouse	b. pastor	c. consent	d. equity
26.	Working these long hours will just …………. you.			
	a. exhaust	b. relieve	c. suspend	d. derive
27.	He is planning to climb the cliff and …………. into the water.			
	a. endorse	b. program	c. plunge	d. prompt
28.	I like wood with a rough …………..			
	a. statue	b. texture	c. parade	d. handle
29.	They were delayed by rough …………..			
	a. entity	b. divorce	c. uniform	d. terrain
30.	You can use your voice to …………. text to your computer.			
	a. dictate	b. trigger	c. damage	d. differ

EXERCISE TWO: Fill-in-the-gap (List B)

1.	They will …………. his hard work.			
	a. behave	b. wander	c. reward	d. invent
2.	They decided to …………. the enemy on the high ground.			
	a. assert	b. depict	c. battle	d. devote
3.	…………. from the aircraft was scattered over a large area.			
	a. handful	b. debris	c. horizon	d. edition
4.	She is a …………. heart surgeon.			
	a. cluster	b. habitat	c. carrier	d. pioneer
5.	It wasn't really my …………. to embarrass him.			
	a. exhibit	b. intent	c. arrival	d. actress
6.	We took steps to …………. their cooperation.			
	a. compel	b. submit	c. resign	d. murder
7.	I had to …………. that I'd overreacted.			
	a. borrow	b. punish	c. exceed	d. concede
8.	One guest at the crowded reception was heard to …………..			
	a. mutter	b. excuse	c. render	d. endure
9.	They woke up early to …………. the fields.			
	a. retain	b. harvest	c. absorb	d. modify
10.	Don't say anything you might …………. later.			
	a. regret	b. signal	c. rescue	d. vanish
11.	They will …………. a fortune from their father.			
	a. inherit	b. survey	c. injure	d. praise
12.	They tried to …………. the cause of the problem.			
	a. desire	b. bounce	c. govern	d. isolate
13.	They had a narrow …………..			
	a. carrier	b. charity	c. escape	d. grocery
14.	The aircraft began to …………..			
	a. convey	b. descend	c. resume	d. admire

15.	Population ………… influences the risk of disease.			
	a. miracle	b. venture	c. density	d. blanket
16.	They found a ………… selling ice cream.			
	a. courage	b. formula	c. gallery	d. vendor
17.	Vitamin C helps your body ………… the iron in your diet.			
	a. occupy	b. confess	c. utilize	d. cancel
18.	He lives in ………… with his neighbours.			
	a. inquiry	b. harmony	c. surgeon	d. counsel
19.	They were in a terrible …………..			
	a. essence	b. pitcher	c. dilemma	d. stretch
20.	He managed to retain his …………..			
	a. hallway	b. monster	c. crystal	d. dignity
21.	There will be penalties if you fail to …………..			
	a. comply	b. debate	c. dismiss	d. possess
22.	He was her friend and ………… until his death.			
	a. warrior	b. mentor	c. curtain	d. playoff
23.	When tourists …………, the town is a very different place.			
	a. invade	b. respect	c. justify	d. squeeze
24.	He experienced the ………… of losing a loved one.			
	a. landing	b. conduct	c. monitor	d. trauma
25.	These people are seeking ………… from persecution.			
	a. vitamin	b. mandate	c. reserve	d. refuge
26.	I ………… you to tell anyone.			
	a. convert	b. pretend	c. forbid	d. witness
27.	Any ………… will need credible plans to win.			
	a. pension	b. nominee	c. shuttle	d. complex
28.	You must obtain a parking ………….			
	a. permit	b. liberty	c. primary	d. pursuit
29.	I am sure that common sense will ………… in the end.			
	a. violate	b. inspire	c. assign	d. prevail
30.	He agreed to be my ………… so that I could join the club.			
	a. racism	b. sponsor	c. gravity	d. servant

EXERCISE THREE: L2-L1 Translation task (List A)

Please translate the underlined word

	الجملة	الترجمة
1.	They were held hostage by armed rebels.	
2.	It is very risky to ride your bike without wearing a helmet.	
3.	Donate the books to a library.	
4.	They cannot publish your name without your consent.	
5.	They were delayed by rough terrain.	
6.	He is planning to climb the cliff and plunge into the water.	
7.	The breeze rippled the water.	
8.	I only got a glimpse of him as we drove by.	
9.	Some plants thrive in shade, while others will not.	
10.	The jury returned a verdict of guilty.	
11.	After their defeat, the soldiers surrendered.	
12.	If you persist with this behaviour, you will be punished.	
13.	You can use your voice to dictate text to your computer.	
14.	If you pull a thread from the sock, it will ruin.	
15.	Be careful not to stumble on the uneven pavement.	
16.	He wore a shirt with a tight-fitting collar.	
17.	Such conditions foster the spread of the disease.	
18.	I like wood with a rough texture.	
19.	We need to exploit this opportunity.	
20.	Working these long hours will just exhaust you.	
21.	She wiped her nose with her sleeve.	
22.	They were unable to continue the race due to fatigue.	
23.	We condemn all acts of terrorism.	
24.	That sweater will shrink if you wash it.	
25.	Insert your credit card into the card machine.	
26.	This vaccine protects against some kinds of the bacteria.	
27.	Would anyone object if we started the meeting now?	
28.	What did you say to provoke him?	
29.	The royal palace is open to the public.	
30.	Don't let this disturb you any more than it already has.	

EXERCISE THREE: L2-L1 Translation task (List B)

Please translate the underlined word

	الجملة	الترجمة
1.	There will be penalties if you fail to comply.	
2.	I forbid you to tell anyone.	
3.	He agreed to be my sponsor so that I could join the club.	
4.	He never recovered from the traumas he suffered during childhood.	
5.	I couldn’t see any way out of the dilemma.	
6.	Population density influences the risk of disease.	
7.	I am sure that common sense will prevail in the end.	
8.	She is a pioneer heart surgeon.	
9.	You must obtain a parking permit.	
10.	A nominee will need credible plans.	
11.	These people are seeking refuge from persecution.	
12.	They will reward his hard work.	
13.	We took steps to compel their cooperation.	
14.	I heard him mutter something under his breath.	
15.	He managed to retain his dignity.	
16.	He ordered the army to invade at dawn.	
17.	They will inherit a fortune from their father.	
18.	They woke up early to harvest the fields.	
19.	It wasn't really my intent to embarrass him.	
20.	He lives in harmony with his neighbours.	
21.	Don't say anything you might regret later.	
22.	They decided to battle the enemy on the high ground.	
23.	I had to concede that I’d overreacted.	
24.	They tried to isolate the cause of the problem.	
25.	They had a narrow escape.	
26.	Vitamin C helps your body utilize the iron in your diet.	
27.	Debris from the aircraft was scattered over a large area.	
28.	He was her friend and mentor until his death.	
29.	The aircraft began to descend.	
30.	They found a vendor selling ice cream.	

EXERCISE FOUR: L2-L1 Translation task (List A)

Please translate the underlined word into English.

الترجمة	الجملة	
	ادخل بطاقة الصراف الآلي.	١.
	تبرع بمالك لإعانة المحتاجين.	٢.
	شجب العالم كله جرائم المحتل.	٣.
	استسلم الجنود بعد الهزيمة النكراء.	٤.
	تأخروا بسبب التضاريس الوعرة.	٥.
	الجدار خشن الملمس.	٦.
	أطلق سراح الرهينة.	٧.
	يجب أن تستغل جميع فرص الحياة.	٨.
	لا يوجد لقاح ضد فيروس نقص المناعة المكتسبة.	٩.
	يمكنك أن تلقن الحديث للحاسوب ويحوله إلى نص مكتوب.	١٠.
	يا له من نسيم عليل!	١١.
	انكمش القميص لأنه غسل بماء ساخن.	١٢.
	شققت كم القميص بعدما علق في عروة الباب وانا أجرى.	١٣.
	تعزز الأجواء الرطبة من انتشار البعوض.	١٤.
	سأقوم بزيارة قصر الملك.	١٥.
	احذر من أن تستفز الحيوانات البرية فقد تهاجمك.	١٦.
	لا يجوز إجراء أي تغيير دون موافقة جميع الشركاء.	١٧.
	لقد شاهدته يغطس من ارتفاع ١٠ أمتار.	١٨.
	تزدهر المجتمعات بالأعمال التطوعية.	١٩.
	تعثر الحصان وأوقع الفارس.	٢٠.
	لقد أصر على وجهة نظره.	٢١.
	العمل لساعات طويلة سيستنزف طاقتك	٢٢.
	ارتدى قميصا بياقة طويلة.	٢٣.
	شعرنا بالإرهاق بسبب الرحلة الطويلة.	٢٤.
	ضع الخيط والإبرة جانبا	٢٥.
	اعذرني! لم اقصد أن أزعجك.	٢٦.
	اعترض اللاعب على قرار الحكم.	٢٧.
	لا تنسى ارتداء الخوذة عند قيادة الدراجة.	٢٨.
	صدر الحكم بالإجماع.	٢٩.
	بالكاد رأيت الطائر فلم أدركه إلا بلمحةٍ واحدة.	٣٠.

EXERCISE FOUR: L2-L1 Translation task (List B)

Please write the English translation of the underlined words.

الترجمة	الجملة	
	ينهى الإسلام عن الغيبة والنميمة.	١.
	لماذا تتمتم في الكلام؟	٢.
	انحدرت المركبة بسرعة على الطريق السريع.	٣.
	لن تستطيع الهروب فالباب مغلق.	٤.
	يعيش في ألفة مع جيرانه.	٥.
	الصين دولة ذات كثافة سكانية عالية.	٦.
	يجب أن يكون لديك كفيل لكي تستخرج قرضا ماليا.	٧.
	غلب فريقنا فريقهم في المسابقة النهائية.	٨.
	يجب اختيار مرشح مناسب.	٩.
	حاول أن تسخر وقت فراغك فيما ينفعك.	١٠.
	أنت بحاجة إلى تصريح دخول.	١١.
	اجتاح الجراد المزرعة.	١٢.
	أين أجد بائع الحلوى؟	١٣.
	كافأ الأب ابنه بمناسبة نجاحه من الجامعة.	١٤.
	تعرضت الأم لصدمة حادة.	١٥.
	عزل السيل الناس عن منازلهم.	١٦.
	انه يعتقد بأن المساعدة في أعمال المنزل تحط من الكرامة.	١٧.
	ذكاء المحقق أرغم اللص على الاعتراف.	١٨.
	تعارك الشرطي مع المجرم.	١٩.
	ندم المتسابق على عدم استعداده جيدا للسباق.	٢٠.
	اقر المتهم بالجرائم المنسوبة له.	٢١.
	ورث عن أبيه ثروة طائلة.	٢٢.
	ابحث عن ملجأ آمن.	٢٣.
	التشرد معضلة كبيرة.	٢٤.
	حطام الطائرة في كل مكان.	٢٥.
	حصد الفلاح محصول هذا العام.	٢٦.
	ستكون العواقب وخيمة إن لم تطع الأوامر.	٢٧.
	أبي قدوتي ومرشدي في الحياة.	٢٨.
	حلمي أن أصبح مستثمر ريادي.	٢٩.
	أحسن النية والتزم بالصبر.	٣٠.

Appendix J: **Participant Information Sheet and Consent Form**

Project title: The Role of Vocabulary Repetition on Vocabulary Retention of Saudi EFL Learners.

What is the project about?

This project aims at examining the impact of repetition practice on vocabulary retention under real classroom conditions.

What does participation involve?

It involves attending and participating in the 104215-3 module (Reading Skills). It also involves participation in a series of tests, questionnaires, and interviews, which will take place during the module. The tests will examine vocabulary proficiency, while the questionnaires and interviews will ask questions about demographic and language learning information, language need, and language use.

Please tick the appropriate boxes	**Yes**	**No**
Taking Part		
I have read and understood the project information given above.	□	□
I have been given the opportunity to ask questions about the project.	□	□
I understand that my taking part is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	□	□
I understand that my decision will not have an impact on the scores of the module; while attendance is a compulsory requirement for the module, I can decline taking part in the tests, questionnaires, and interviews.	□	□
I agree to take part in the project.	□	□
Use of the information I provide for this project only		
I understand my personal details such as name, email address and phone number will not be revealed to people outside the project.	□	□
I understand that my words may be quoted in publications, reports, web pages, and other research outputs.	□	□
I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand that other genuine researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□

________________________ ________________________ ________

Name of participant [printed] Signature Date

________________________ ________________________ ________

Researcher Date

Contact details: [Email: ________Mobile KSA: Mobile UK: