ADVANCED LEARNERS' RETENTION OF LOW-FREQUENCY VOCABULARY: THE EFFECTS OF TASK TYPE

KIYMET MERVE CELEN

BOĞAZİÇİ UNIVERSITY

2023

ADVANCED LEARNERS' RETENTION OF LOW-FREQUENCY VOCABULARY: THE EFFECTS OF TASK TYPE

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

English Language Education

by

Kıymet Merve Celen

Boğaziçi University

2023

Advanced Learners' Retention of Low-Frequency Vocabulary:

The Effects of Task Type

The thesis of Kıymet Merve Celen

has been approved by:

Assist. Prof. Şebnem Yalçın _____
(Thesis Advisor)

Prof. Gülcan Erçetin _____

Assoc. Prof. Senem Yıldız _____

Assist. Prof. Derya Altınmakas _____
(External Member)

Assist. Prof. Mustafa Polat _____
(External Member)

July 2023

DECLARATION OF ORIGINALITY

I, Kıymet Merve Celen, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....

ABSTRACT

Advanced Learners' Retention of Low-Frequency Vocabulary:

The Effects of Task Type

This study aimed to investigate task type effects on advanced level English language learners' early and long-term retention of low-frequency vocabulary. Two reading+replacing tasks which varied in their involvement load indexes (Laufer & Hulstijn, 2001) were created (i.e., LowText = index of 2 vs. HighText = index of 1). Advanced level university students were randomly assigned to one of the tasks which required them to read a text and replace the ten words embedded in it with their synonyms from the opposite word-frequency family (i.e., low- or high-frequency) provided in a word list. Immediate and delayed posttests measured the retention of the target words two weeks apart. The results showed that the tasks with different involvement load indexes did not lead to significantly different retention of the target words, neither in the short term nor in the long term. Participants' knowledge of the target words did not show a significant increase after completing the tasks, and their long-term retention of the target words was significantly lower than their previous knowledge. The results are discussed in relation to target word characteristics, task design, participant performance, and tests used to measure target word knowledge.

ÖZET

İleri Seviye İngilizce Öğrenenlerin Düşük Sıklığa Sahip Kelime Öğrenimleri: Görev Türünün Etkileri

Bu çalışma, görev türünün düşük sıklıktaki kelimelerin ileri seviye İngilizce yeterliliğine sahip öğrenciler tarafından kısa ve uzun vadede öğrenimi üzerindeki etkisini incelemeyi amaçlamaktadır. Bu amaçla, ilgi yükü endeksleri (Laufer & Hulstijn, 2001) açısından farklılık gösteren iki okuma+kelime değiştirme görevi dizayn edilmiştir (DüşükMetin = endeks 2 ve YüksekMetin = endeks 1). İleri seviye İngilizce yeterliliğine sahip üniversite öğrencileri, kendilerinden bir metin okuyup bu metnin içindeki kelimeleri, sunulan listede bulunan zıt sıklığa sahip (düşük ya da yüksek sıklık) eş anlamlı kelimelerle değiştirmelerinin istendiği gruplardan birine rastgele atanmıştır. Kelimelerin öğrenimi iki hafta aralıkla yapılan son testlerle ölçülmüştür. Sonuçlar, farklı ilgi yükü endekslerine sahip görevlerin hedef kelimelerin öğreniminde hem ilk ve hem de gecikmeli son testlerde anlamlı bir fark yaratmadığını göstermiştir. Görevlerini tamamladıktan sonra katılımcıların hedef kelime bilgilerinde anlamlı bir artış görülmemiş; katılımcıların uzun vadedeki kelime öğrenimleri kelime önbilgilerinden anlamlı bir şekilde düşük bulunmuştur. Sonuçlar hedef kelime özellikleri, görev dizaynı, katılımcı performansı ve hedef kelime bilgisini ölçmekte kullanılan testler açısından tartışılmaktadır.

CURRICULUM VITAE

NAME: Kıymet Merve Celen

DEGREES AWARDED

PhD in English Language Education, 2023, Boğaziçi University
MA in English Language Education, 2016, Boğaziçi University
BA in Foreign Language Education, 2012, Boğaziçi University

AREAS OF SPECIAL INTEREST

L2 vocabulary learning/teaching, lexical sophistication, lexical richness, teacher education, and program evaluation

PROFESSIONAL EXPERIENCE

English Language Instructor
School of Foreign Languages, Istanbul University, 2019 - present

English Teacher
Istanbul University Itrî Fine Arts High School, 2021 - 2022

Research Assistant
Department of Foreign Language Education, Yıldız Technical University, 2013 - 2019

English Language Instructor
Department/School of Foreign Languages, Dumlupınar University, 2012 - 2013

GRANTS

Teaching Excellence and Achievement Program for Turkish Pre-service Teachers, 2011
Bureau of Educational and Cultural Affairs of the U.S. Department of State

PUBLICATIONS

Celen, K. M., & Yalçın, Ş. (2021). The effects of vocabulary resource use on lexical richness in L2 writing. *Millî Eğitim, 50*(230), 1039-1058.

Celen, K. M., & Akcan, S. (2017). Evaluation of an ELT practicum programme from the perspectives of supervisors, student teachers and graduates. *Journal of Teacher Education and Educators, 6*(3), 251-274.

CONFERENCE PRESENTATIONS

Celen, K. M., & Yalçın, Ş. (2021, August). *Designing tasks for advanced L2 vocabulary learning through the manipulation of lexical sophistication of reading texts.* 19th World Congress of Applied Linguistics on the Dynamics of Language, Communication and Culture in a Changing World, Groningen, the Netherlands (Online).

Celen, K. M., & Yalçın, Ş. (2019, September). *Lexical variation and sophistication in L2 writing: Tracking word changes in an essay revision task.* Topics in Applied Linguistics: Classroom-oriented Research, Opole, Poland.

Celen, K. M., & Yalçın, Ş. (2018, September). *Lexical sophistication and variation in L2 writing: Exploring the effects of dictionary and thesaurus use.* The 51st Annual Meeting of the British Association of Applied Linguistics (BAAL), York, UK.

Celen, K. M., & Akcan, S. (2016, August). *Evaluation of an English language teacher education practicum: Insights from supervisors, student teachers, and graduates*. International Symposium on New Issues in Teacher Education (ISNITE), Savonlinna, Finland.

Celen, K. M. (2016, May). *Program evaluation of an English language teacher education practicum: Insights from student teachers and graduates*. The 9th International ELT Research Conference. Çanakkale, Türkiye.

Celen, K. M., & Bayyurt, Y. (2015, October). *Brand naming practice from a linguistic perspective: A case in Turkey*. The 21st Conference of the International Association for World Englishes. Istanbul, Türkiye. (Poster presentation)

Celen, K. M. (2014, October). *On the nature of and motives for topic shift in Turkish*. International Conference for Academic Disciplines. Rome, Italy.

Celen, K. M., & Yüksel, H. G. (2014, May). *Vocabulary size and collocational knowledge: Are they related?* The 8th International ELT Research Conference. Çanakkale, Türkiye.

TRAININGS & CERTIFICATES

Understanding English Dictionaries, 2019
Coventry University, The Alan Turing Institute and Macmillan Education
Online course by Coventry University

Fall 2018 Five-Week Massive Open Online Course
Professional Development for Teacher Trainers, 2018
Online course by Arizona State University

Certificate in Linguistics, 2012
Department of Western Languages and Literatures of the Faculty of Arts and Sciences, Boğaziçi University

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Assist. Prof. Şebnem Yalçın, and the committee members, Prof. Gülcan Erçetin, Assoc. Prof. Senem Yıldız, Assist. Prof. Derya Altınmakas, and Assist. Prof. Mustafa Polat, for their valuable feedback and guidance. I would also like to thank the instructors who allowed me to visit their classes to explain my study to recruit participants. I am grateful to my second rater, whose hard work made various analyses possible, and also to my friends who completed my tasks and provided feedback. Many thanks go to the participants who made time to join the sessions.

Canım annemin desteği bana hep güç verdi.

This PhD study was funded by Boğaziçi University Research Fund grant number 15742D.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: BACKGROUND 7

2.1 The Involvement Load Hypothesis 7

2.2 Vocabulary and reading 27

2.3 Reformulation, revision, and editing 33

2.4 Incidental/Intentional learning and focus on forms 35

2.5 Task interactiveness and experimental tasks 36

2.6 Word frequency and word families 38

CHAPTER 3: METHODOLOGY 41

3.1 Participants 42

3.2 Instruments 45

3.3 Procedure 55

3.4 Data analysis 56

CHAPTER 4: RESULTS 60

4.1 Research question 1 60

4.2 Research question 2 61

4.3 Research question 3 62

CHAPTER 5: DISCUSSION 96

5.1 Task properties 100

5.2 Testing instruments 103

CHAPTER 6: CONCLUSION 111

6.1 Pedagogical implications 112

6.2 Limitations and suggestions for further research 113

APPENDIX A: TASK A (LOWTEXT) 115

APPENDIX B: TASK B (HIGHTEXT) 121

APPENDIX C: THE MINI-DICTIONARY 127

APPENDIX D: READING COMPREHENSION QUESTIONS 128

APPENDIX E: THE SURVEY OF TASK INTERACTIVENESS 130

APPENDIX F: THE TEST OF PREKNOWLEDGE 131

APPENDIX G: IMMEDIATE AND DELAYED POSTTESTS: THE VKS 132

APPENDIX H: THE TEST OF FORM RECALL 138

APPENDIX I: ETHICS COMMITTEE APPROVAL 141

REFERENCES 142

LIST OF TABLES

Table 1. Task-induced Involvement Load (Laufer & Hulstijn, 2001, p. 18) …….….9

Table 2. Materials for Tasks A and B ……..…....…46

Table 3. Involvement Loads of the Reading+Replacing Tasks ….…….....….48

Table 4. Target Words, Their Counterparts, and Extra Words ……..…...….49

Table 5. VKS Elicitation Scale (Paribakht & Wesche, 1997, p. 180) …...……..….52

Table 6. Orders of the Posttests ……..……...54

Table 7. Descriptive Statistics for Target Word Knowledge …….……...58

Table 8. Number of Participants Recalling the Target Words ……..…….61

Table 9. Word Replacement Activity Number of Correct/Wrong Answers …….....63

Table 10. Participants with Highest Scores in at Least Four Score Categories ….....77

Table 11. Participants with Lowest Scores in at Least Four Score Categories ….....78

Table 12. Participant Profiles for the Highest and Lowest Scores in the Immediate Posttest (Dichotomous) ……...……....79

Table 13. Participant Profiles for the Highest and Lowest Scores in the Delayed Posttest (Dichotomous) …….……..80

Table 14. Participant #1's Pre- to Posttest Target Word Knowledge ……..…...104

Table 15. Participant #31's Pre- to Posttest Target Word Knowledge ...…...…….107

Table 16. Participant Responses Treated as Incorrect for Words with Multiple Meanings……...…….108

LIST OF FIGURES

Figure 1. Interactiveness (Bachman & Palmer, 1996, p. 26) ……37

Figure 2. Participant age (top), major (middle), and year in the degree program (bottom)……..…….43

Figure 3. Pass scores for those completing the preparatory English program ….....44

Figure 4. Exam results received by those who did not attend the preparatory program……...……...……..44

Figure 5. English language learning history ……..……...45

Figure 6. VKS score meanings (Paribakht & Wesche, 1997, p. 181) ..…...…….57

Figure 7. Group A's (left) and Group B's (right) word replacement task performance……..…….……..62

Figure 8. Reading comprehension performance by groups ……..…64

Figure 9. Participant ratings for task interactiveness ……..……..65

Figure 10. Group A's preknowledge of target words ……...……...67

Figure 11. Group B's preknowledge of target words ..…….68

Figure 12. Posttest performance for Group A ……...……...…….69

Figure 13. Posttest performance for Group B ……...……70

Figure 14. Group A's (top) and Group B's (bottom)VKS score percentages for *bungle*……..……73

Figure 15. Group A's (top) and Group B's (bottom) VKS score percentages for *crass*……..……...74

Figure 16. Group A's (top) and Group B's (bottom) VKS score percentages for *detractor*……..……76

Figure 17. *Ramification:* Changes in posttest scores (top) and posttest scores (bottom)……..…….82

Figure 18. *Bungle:* Changes in posttest scores (top) and posttest scores (bottom)……...…83

Figure 19. *Muddled:* Changes in posttest scores (top) and posttest scores (bottom)……..….84

Figure 20. *Confrontational:* Changes in posttest scores (top) and posttest scores (bottom) ……..……85

Figure 21. *Lament:* Changes in posttest scores (top) and posttest scores (bottom)……...86

Figure 22. *Crass:* Changes in posttest scores (top) and posttest scores (bottom)…….……..87

Figure 23. *Premise:* Changes in posttest scores (top) and posttest scores (bottom)……...88

Figure 24. *Desecrate*: Changes in posttest scores (top) and posttest scores (bottom)……...89

Figure 25. *Dissection:* Changes in posttest scores (top) and posttest scores (bottom)……...90

Figure 26. *Detractor:* Changes in posttest scores (top) and posttest scores (bottom)……...91

Figure 27. Total number of cases for immediate*delayed posttest scores …..…….92

Figure 28. Number of participants with progress and decline in scores across posttests……...…..……93

Figure 29. Magnitude of progress and decline per each word across posttests ..…..94

Figure 30. The electronic word replacement task sample ……..…102

Figure 31. The electronic-to-print word replacement task sample ……....103

Figure 32. Target words marked ‘known’ initially but ‘unknown’ later ……...106

ABBREVIATIONS

BNC	British National Corpus
ILH	Involvement Load Hypothesis
L1	First Language
L2	Second Language
NFL7	Nuclear Family List 7
TFA	Technique Feature Analysis
VKS	Vocabulary Knowledge Scale

CHAPTER 1

INTRODUCTION

Research on the teaching and learning of second language (L2) vocabulary seems to be shaped by the very nature of language: the limited (or countable) numbers of grammatical rules and incomparably larger amounts of lexical units to master. The immensity of the amount of the words to learn has been called a "challenge" (Schmitt, 2008, p. 331); research-informed conclusions have been made to point out that "some words are much more useful than others" (Nation, 2001a, p. 9); and suggestions have been made for targeting more realistic numbers of words in the L2 classroom (Schmitt, 2000) and incorporating certain clearly stated goals in syllabuses and lessons (Milton, 2009). According to Sinclair and Renouf (1988), for all English language learners, the most frequent words along with their various uses need to be "the main focus of study" (p. 148). Naturally, high-frequency words deserve the teaching time allocated to them due to "their frequency, coverage, and range" (Nation, 2001a, p. 16) and direct instruction may be preferred for teaching the most frequent words (Grabe, 2009; Horst, Cobb, & Meara, 1998). The uses of direct teaching, however, does not seem to be limited to high-frequency words. Coxhead (2000), for instance, suggested that the direct teaching of academic words in her new Academic Word List is a worthwhile endeavor. Low-frequency words, on the other hand, do not require direct teaching, but can be handled more through strategy training (Laufer & Nation, 1999). The learning of less frequent words also relies more on individual investment and it appears that good learners "create their own opportunities for exposure and repetition" (Milton, 2009, p. 242). No matter how practical it is for teaching purposes to divide words into high and low frequency

groups, it is a synthetic one as "some words are not frequent, but they are not infrequent either, and in specialized fields may be quite common" (Xue & Nation, 1984, p. 215). In addition, frequency lists need to be considered "more as a useful indication rather than a prescription" (Vilkaitė-Lozdienė & Schmitt, 2020, p. 88).

In addition to the very characteristics of the words themselves, what serves a good means for vocabulary acquisition in an L2 has attracted considerable research attention. Paribahkt and Wesche (1997) suggested that focused instruction is a useful alternative "when the learning period is limited and specific vocabulary outcomes are sought" (p. 197). Similarly, Laufer (2003) argued that word-focused activities could prove more beneficial and more time-saving when contrasted with reading. Schmitt (2008), likewise, contested the idea that "an adequate lexis will simply be 'picked up' from exposure to language tasks focusing either on other linguistic aspects or on communication" (p. 333).

One issue that emerges is how the learning as well as the active use of low-frequency words can still be relevant for L2 learners, especially for those who have reached advanced levels of proficiency. In a study which provided valuable insight into this issue, Laufer (1991) underscored the importance of raising standards higher when it comes to advanced learners' productive use of L2 vocabulary:

> If the tendency of L2 learners is to remain at the threshold level, it is the task of the teacher to elicit the above-threshold vocabulary, which is precisely the vocabulary that learners try to avoid. Whatever form this elicitation might take (asking for words with different shades of meaning, reformulating sentences, gap filling, translation from L1 to L2, etc.) its goal is to activate the vocabulary which may otherwise remain at the passive end of the vocabulary knowledge continuum. (p. 446)

Laufer's (1991) study approached one end of the advanced learners' vocabulary knowledge, more specifically their use. To take this further would be to sustain advanced learners' continuous learning of low-frequency words, which would then

best be put to active use with the help of strategies put forward by Laufer. The idea of targeting such an achievement in L2 vocabulary learning appears to be neither quirky nor ambitious. Referring to a review of vocabulary learning literature, Bardel (2016) pointed to a lack of a theoretically-sound hindrance to a nativelike attainment in L2 vocabulary. Bardel also suggested that "a learner should be able to develop specialized vocabulary to a level that might even supersede the 'average' native speaker's vocabulary" (p. 101). This is where the present study situates the driving forces behind it: advanced learners, low-frequency words, and identifying tasks which can cater for both.

It becomes important to discuss what advanced proficiency is at this point. Research on advanced L2 abilities, however, does not appear to have a long history. Not so long ago, Byrnes (2004, 2006) pointed to a lack of understanding of the characteristics of advanced L2 abilities. Bardel (2016) likewise pointed to a lack of uniformity found in various categories of advancedness found in the literature, and according to her definition, an advanced user or learner is "a person whose L2 is close to that of a native speaker, but whose non-native usage is perceivable in normal oral or written production" (p. 75). Bardel further underscored the relatively different status of vocabulary when compared to other aspects of a language, in a discussion of native abilities:

> When it comes to vocabulary it is important to be aware of the fact that this is an area of language that is – maybe more than any other – characterized by variation among L1 speakers as well, depending on various social, situational and cognitive factors such as education and literacy, context, memory, concentration, level of formality, etc. Vocabulary knowledge varies not only among L2 learners at different levels of proficiency, but also among native speakers. (p. 76)

Similarly, Milton (2009) explained how native speakers (e.g., undergraduates starting tertiary education, post-graduates, and faculty members) even differ from

each other with respect to their vocabulary knowledge and suggested that it might indeed be better to target a level of vocabulary knowledge that is comparable to the learner's L1 (first language) instead of aiming for a definite vocabulary size which might not be reached by native speakers themselves on various occasions.

Referring to what professionals working with advanced language learners suggest, Byrnes (2006) noted that it is not essentially a compliance (or lack thereof) with the rules of grammar when teaching and learning from an advanced proficiency perspective is concerned; rather, it is about "making choices and the capacity to make those choices in a meaningful—that is, culturally and situationally conscious—fashion, including deliberate and now meaningful violations of 'rules' and 'fixed norms'" (p. 5). As can be understood, advanced learners, in this respect, differ from learners with lower levels of proficiency in terms of the target behaviors expected of them: They are still language learners but are required to function at a larger arena of language use. Although the American college foreign language programs have mainly conceptualized advanced second language abilities by drawing on those defined by ACTFL[1] (Byrnes, 2004), it should be noted that different contexts and institutions have adopted other descriptors of performance and tools for assessing test-takers' performance and the resulting description of their language abilities. Byrnes (2004) further recommended four principles for teaching advanced level learners: (1) a focus on cognitive aspects of learning, (2) explicit teaching, (3) use of modelling, coaching, and scaffolding, and (4) task-based teaching.

Taking the above discussion into consideration and following Laufer and Hulstijn's (2001) call for expanding the scope of their Involvement Load Hypothesis (ILH) by way of suggesting additional components, this study aimed to investigate

[1] American Council on the Teaching of Foreign Languages

the effectiveness of two tasks on advanced leaners' retention of low-frequency L2 words.

Focusing on young adults receiving education at a state university where the medium of instruction is English, this study aimed to look for ways to sustain the development of lexis in advanced stages of the language learning process by using a task which came in the form of an after-reading task, but which was believed to hint at a lexically richer language production expected of learners of more advanced proficiency levels. In other words, this study primarily dealt with how low-frequency words can be taught and/or learned and, as an ancillary aim, it drew attention to the use of low-frequency words in language production by focusing on how the same base text *felt* with and without them.

Naturally, in instructed language learning settings, there is an end to the language learning process via in-class instruction. Hopefully, by the end of this period, institutional, regional, or country-level objectives are met. The process what we call as language instruction usually welcomes learners with little or no knowledge of the target language at the lowest and graduates them at various exit levels. Usually, again, these exit levels are below advanced levels of proficiency. This is normal because not all learners wish to attain such advanced levels, and they might stay and function well as, say, intermediate level speakers. However, there are those who, for personal or professional reasons, need to add more to their language repertoires. Instructed or otherwise, such learners will continue to learn the aspects of the target language which was not-so-essential during their initial language learning process. These aspects might include less common or delicate grammatical structures, more sharpened four skills, and a deeper knowledge of familiar words and more words in general. To continue with the latter aspect, such a need might be for

(a) a deeper or more precise expression or (b) stylistic reasons for productive use of the language. From a receptive use perspective, a need for knowing more words and knowing more about them will provide learners with immediate, automatic understanding. For all these reasons, for advanced language learners' continued development of lexis, (a) strategy training for low-frequency vocabulary learning and (b) fostering an attitude in favor of rich lexis seem to be crucial. All the rightful attention that learners with lower-proficiency elicit should not prevent language professionals from encouraging advanced language learners to achieve their very best. These learners, with correct guidance, be it in the form of materials, strategies, and even expectations, can achieve even more thanks to the knowledge and skills they have already gained.

To researcher's knowledge, ILH has not been researched by comparing two tasks which required the same procedure, i.e., word replacement, but in the opposite directions. Therefore, this study will hopefully contribute to the body of research focusing on ILH, and naturally on task effectiveness. With its emphasis on advanced learners' continuing development of L2 lexis and materials which can encourage this, the present study is expected to offer a different insight into the field of vocabulary teaching and learning inside and outside the classroom as well as into L2 material design.

CHAPTER 2

BACKGROUND

This section briefly reviews the core concepts this study incorporates in its design. Explanations of concepts as well as examples from the literature are provided to better explain the decisions behind the study design.

2.1 The Involvement Load Hypothesis

Laufer and Hulstijn's (2001) Involvement Load Hypothesis draws on two papers on processing and memory (Craik & Lockhart, 1972; Craik & Tulving, 1975) and it is an embodiment of a "need to translate and operationalize [depth of processing and elaboration] in terms of L2 vocabulary learning tasks" (Hulstijn & Laufer, 2001, p. 543). In their paper introducing the construct of task-induced involvement, Laufer and Hulstijn (2001) refer to the psychological literature to explain how *elaboration* has been considered to facilitate the retention of words and they further acknowledge the importance of *motivation* in learning an L2. Accepting the importance of these two constructs for vocabulary learning as well, they contend that theory and research on concepts of cognition of these kinds within the area of L2 vocabulary learning has not reached the level the area L2 grammar learning has attained. Motivated by this and in an effort to suggest a measure of the depth of processing required by a given task, the authors introduce *involvement*, a construct made of up three dimensions: *need*, *search*, and *evaluation*. *Need* is proposed as the motivational dimension, unlike *search* and *evaluation* which constitute the cognitive dimensions of the construct. It can be moderate or strong, depending on the source causing it to emerge. When caused by an outside factor (i.e., when a teacher asks a learner to use a word in a

sentence), *need* is considered to be moderate; however, when the learner herself feels the need to use or understand the meaning of a word, *need* is considered to be strong. *Search* is described as the process where the meaning of an L2 word or its form (when the L1 word is known) unknown to a learner is sought by referring to a dictionary or other sources of information such as a teacher. Laufer (2020) later mentioned that a recommendation was made for separate categories of moderate and strong search, which was not the case initially. *Evaluation* is considered to occur in situations such as when one compares a particular word with others (or one of the meanings of a word with the other meanings it has) or when a learner decides on the suitability, in terms of form and meaning, of a word within the context of others words. Just like *need*, *evaluation* can be moderate or strong. When it involves making a differentiation among a group of words or among the multiple meanings of a word in a certain context, evaluation is considered to be moderate, but when the participant needs to make judgments about other words which will go together with the new word in a new sentence or a piece of writing, strong *evaluation* is expected to occur. The way Laufer and Hulstijn's (2001) *evaluation* functions in this study has a close resemblance to what Paribakht and Wesche (1997) refer to as interpretation; more specifically, one of the examples of it suggested as "[u]nderstanding the meanings and grammatical functions of the target word in the text (i.e., in a given context) and recognizing words or phrases that could be substituted in the text" (p. 184). Table 1 illustrates Laufer and Hulstijn's list of tasks that vary in their extent to which *need*, *search*, and *evaluation* factors are included:

Table 1. Task-induced Involvement Load (Laufer & Hulstijn, 2001, p. 18)

Task	Status of target words	Need	Search	Evaluation
1. Reading and comprehension questions	Glossed in text but irrelevant to task	-	-	-
2. Reading and comprehension questions	Glossed in text and relevant to task	+	-	-
3. Reading and comprehension questions	Not glossed but relevant to task	+	+	-/+ (depending on word and context)
4. Reading and comprehension questions and filling gaps	Relevant to reading comprehension. Listed with glosses at the end of the text	+	-	+
5. Writing original sentences	Listed with glosses	+	-	++
6. Writing a composition	Concepts selected by the teacher (and provided in L1). The L2 learner-writer must look up the L2 form	+	+	++
7. Writing a composition	Concepts selected (and looked up) by L2 learner-writer	++	+	++

Note. - = absent. + = moderately present. ++ = strongly present.

Overall, Laufer and Hulstijn (2001) suggest that "[t]he combination of factors with their degrees of prominence constitutes involvement load" (p.17) and that tasks having higher involvement load indexes (as calculated by the summation of their relative index points), when other factors are kept constant, will better predict the retention of vocabulary in comparison to tasks having lower involvement load. Hulstijn and Laufer (2001) went on to test ILH and conducted two experiments with advanced learners to investigate the retention of 10 low-frequency words and expressions by using three tasks differing in their involvement loads. The tasks used were (1) reading comprehension with marginal glosses (need: +, search: -,

evaluation: -), (2) reading comprehension plus fill in (need: +, search: -, evaluation: +), and (3) writing a composition and incorporating the target words (need: +, search: -, evaluation: ++). For these tasks which received involvement indexes of 1, 2, and 3 respectively and which were completed under incidental learning conditions, the authors anticipated a gradual increase in retention from the first task through the third one. Results coming from both groups confirmed the prediction that the writing task would elicit significantly higher retention scores. Yet, significantly better retention as a result of undertaking the second task in comparison to the first task was observed in one of the groups, not the other.

In a study replicating, in concept, Hulstijn and Laufer (2001), Keating (2008) aimed to compare the retention of eight pseudowords in three different tasks varying in their involvement loads. Beginner level L2 learners of Spanish completed three tasks which varied in the degrees of evaluation they required (i.e., Task 1: absent, Task 2: moderate, and Task 3: strong) but identical in terms of their load concerning the need and search dimensions. Measures of participants' performance were based on scores obtained from active and passive recall tests and partial knowledge of the target words was also given points. The results showed that, in terms of passive recall, although Task 2 and Task 3 produced significantly better retention scores than Task 1, Task 3 was not found to be better than Task 2. Keating offered two possible reasons, namely the selection of the *post hoc* test and a different type of task used for Task 3, to explain the different results found in this study vis-à-vis Hulstijn and Laufer (2001). As for active word recall, the results of the immediate posttest were in complete accordance with ILH, with significant increases in retention scores paralleling increased levels of task-induced involvement. In delayed posttest results, however, only Task 2 was found to elicit significantly better performance than Task

1. Finally, when time on task was controlled, the superiority of Task 2 and Task 3 over Task 1 disappeared. Although the participants had been informed of the two upcoming posttests, Keating listed a number of reasons why the learning taking place was indeed far from intentional. Overall, this study provided further evidence for task-induced involvement in learning L2 vocabulary.

In a within-subjects design, Folse (2006) tested the effectiveness of three types of written exercises on the retention of 15 target words. With the help of a minidictionary provided, ESL learners in three different proficiency groups completed one fill-in-the-blank exercise, three fill-in-the-blank exercises, and an original-sentence-writing exercise, each of which included five of the target words. The idea of including three fill-in-the-blank exercises came from the result of Folse's pilot study which revealed that the time taken to write sentences almost tripled the time spent for a completion exercise. For this reason, this study also adopted a time on task perspective in addition to differences in exercise type. Target words included only verbs in order to eliminate any effect the parts of speech could have on the results. The results showed that participants completing three fill-in-the-blank exercises gained significantly higher retention scores when compared to other types of exercises and it differed from other tasks with statistical significance. When time on task was kept constant, completing three fill-in-the-blank exercises, again, led to significantly better learning than writing original sentences. Folse concluded that "doing multiple target word retrievals in an exercise, no matter how superficial the exercise may seem, is a stronger and more facilitative factor in L2 vocabulary learning than the purported deeper processing or involvement load that writing original sentences with new L2 vocabulary may offer" (p. 287). By adding that writing original sentences is also a task which requires considerable amount of

teacher as well as student time, Folse seemed to disfavor its use rather too quickly—when no information regarding the long term retention of target words was available.

In a study (partially) replicating Folse (2006), Jahangiri and Abilipour (2014) looked into the effects of exercise type and collaboration on the retention of eight target verbs, in an attempt also to put ILH to the test. Intermediate level participants, individually and in pairs, completed two fill-in-the-blank exercises and a sentence writing exercise for each target word within the same amount of time allocated and by using the minidictionaries provided for each one of the four tasks they completed. For the short-term and long-term retention of target vocabulary, no significant differences were found for the effects of exercise type, which ran contrary to the premises of ILH, or for collaboration. The only significant difference was found for the interaction between the two for the long-term retention of target words, with the writing of sentences benefiting from collaborative work and fill-in-the-blanks from individual work. One of the reasons speculated in explaining the failure of collaboration to show positive effects on word retention in the short-term was related to vocabulary as a target language form not being amenable to collaboration. The authors suggested that future research focus on the short- and long-term effects of individual and collaborative work by using different exercise types.

More recently, Hu and Nassaji (2016) compared ILH with Technique Feature Analysis (TFA) (Nation & Webb, as cited in Hu & Nassaji, 2016) in terms of how successfully they determine task effectiveness in L2 vocabulary learning. According to the authors' descriptions, TFA is a more detailed framework than ILH, with its five components adding up to 18 criteria in total. Low-intermediate learners completed four tasks[2] which differed in their respective rankings as predicted by both

[2] Reading a text and rewording the sentences, one of the tasks used in this study as suggested by Nation and Webb, has some resemblance to the text reformulation task used in the present study.

frameworks, with 14 target words embedded in them. When task performances were compared with each other, four of the comparisons yielded significant differences, yet the predictions made by ILH and TFA could not explain three of these differences. The only comparison showing significant difference was in line with TFA, but not with ILH. TFA was further found to better predict vocabulary gains from pretest to posttest than ILH and with statistical significance. Highest retention score was associated with a task involving production, which was highest in terms of evaluation load across the tasks. It is important to note, however, that the calculations of task scores in this study and the very classification of tasks based on them (high and low) can be criticized on many grounds (and the results as well). Additionally, as Hu and Nassaji point out, the tasks used had very similar involvement load indexes (indeed the same can be argued for TFA scores as well), and this might have prevented the authors from distinguishing the tasks from each other with due precision and speculating on their effectiveness accordingly. Laufer (2020) also pointed out the problematic aspects of the paper.

Huang, Willson, and Eslami (2012) conducted a meta-analysis of research focusing on the effects that output tasks (their involvement loads, more specifically) have on learning L2 vocabulary incidentally. Twelve studies targeted for the meta-analysis were scrutinized in relation to (a) the quality of the study design, (b) type of tasks, (c) time spent on task, (d) text type, and (e) text-target word ratios as individual parameters. The results revealed that when studies were medium or high quality in terms of design, the chances for finding statistical significance concerning the learning of target words were significantly higher. Significant differences were reported for different tasks, with the mean effect sizes being highest for participants completing a combination of tasks, but decreasing, in order, for participants

completing an essay writing task, sentence writing task and fill-in-the-blank activities. Analyses of time spent on tasks, likewise, produced significant relationships. It was found that participants spending more time on task also learned more words. As for text genre, significant differences were detected; participants engaging in a combination of different texts received the highest mean effect sizes, but a decrease was observed, in order, for those reading expository and narrative texts, though, due to the limited sample sizes included for the text types, the authors remained cautious about this finding. Unlike the variables mentioned so far, text-target word ratios did not yield statistically significant differences, yet the mean effect sizes found for participants reading texts with the number of target words amounting to fewer than 2% of the words in the texts were larger than those reading texts with this figure set between 2% and 5%. It was concluded that this meta-analysis provided positive evidence for ILH.

In the Turkish context Karalık and Merç (2016) compared fill-in with glossary, fill-in by searching, retelling with glossary, and retelling by searching tasks which had involvement load levels of 2, 3, 3, and 4 respectively. These tasks were based on two reading texts which included a different set of 10 target words each and were given to participants at two different times. In a between-subjects design, participants with a minimum of upper-intermediate level language proficiency completed one of the four tasks. Participants' performance in the immediate and delayed posttests was compared for tasks with different and similar involvement load levels. A comparison including all four tasks indicated that tasks with higher involvement load levels led to better performance both in the immediate and delayed posttests. However, significant differences between the tasks were found only in the delayed posttests which pointed to the superiority of retelling by searching task over

fill-in by searching and fill-in with glossary. For the tasks having the same involvement load levels (i.e., fill-in by searching and retelling with glossary), no significant differences were found based on immediate and delayed posttest results. The authors concluded that their findings were highly compatible with ILH and underscored the reliability of involvement load levels across the texts and target words which were not identical in all four task conditions. They further pointed to better results elicited out of the tasks when a search component (rather than provision of a glossary) was present or when search and strong evaluation co-occurred.

In their meta-analysis, Yanagisawa and Webb (2021) examined 42 previous studies which focused on how successfully ILH predicted learning, and they also looked into the extent to which the three components of the hypothesis (i.e., need, search, and evaluation) and other variables (i.e., time spent on task, frequency, type of vocabulary knowledge, and proficiency) affected the learning of L2 vocabulary incidentally. They found that ILH significantly predicted learning and could explain 15.0% of the effect size (ES) variance for immediate and 5.1% for delayed posttests. The analyses of immediate posttests revealed that need and evaluation components significantly predicted learning gains. For a task with a moderate need component, estimated ESs were reported to be 30.2% larger than a task with no need. For tasks with moderate and strong evaluation, ESs estimated were 13.9% and 22.6% larger than tasks without an evaluation component. Search, as a task component, however, did not predict ESs significantly. The evaluation component had the largest impact on learning, after which need was observed. However, search did not have an effect on learning. Yanagisawa and Webb also calculated the extent of the effect of each component by controlling for the effects of the other components. It was seen that retention of 15.4% of unknown vocabulary was possible when none of the

components was present. Only need accounted for a 20.0% increase in learning. Moreover, the impact of strong evaluation nearly doubled that of moderate evaluation, and these two were found to be significantly different. As for the delayed posttests, similar results were reported. That is, learning was predicted by need and evaluation with significance, but not search. It was seen that retention of 12.7% of unknown vocabulary was possible when none of the components was present. An increase of 12.6% in retention was possible with the addition of moderate need. 21.7% and 27.9% increases in retention were observed when moderate evaluation and strong evaluation were present, respectively, in comparison to the conditions lacking need and search. For time on task, Yanagisawa and Webb reported "that longer learning conditions do not necessarily lead to greater learning gains, but rather learning conditions with larger [involvement loads] tend to take longer, and [involvement load] contributes to learning more than time on task" (p. 510). Frequency, defined as the number of times participants saw or used a target word, yielded a trend. When the involvement load effects were controlled for, an estimated 8.3% increase in learning was observed with every increase (i.e., 1) in frequency in the immediate posttest; however, frequency effects vanished in the delayed posttests. As for the aspect of vocabulary knowledge tested, analyses of both immediate and delayed posttest results revealed significant main effects. With the effects of involvement load controlled for, the largest gains were recorded, in order, for form, form-meaning recognition, form-meaning recall, the VKS, and use (the last two of which changed order in the delayed posttests). It was argued that the involvement load showed "weaker effects on the development of use knowledge or the VKS's developmental stages of word knowledge compared to the development of form and form-meaning knowledge" (p. 512). Finally, for language proficiency, no main

effects were found for both immediate and delayed posttests. Yanagisawa and Webb concluded that ILH found moderate support in their analyses and the correlation between involvement load and vocabulary gains was clear. However, they argued that ILH's predictive ability was not very high. The authors also pointed to the possibility of confusion concerning the definition/operationalization of the components of ILH such as need (when, for example, a learner's motivation to use a word coincides with a teacher-prompted use of it) and search.

Kim (2008a) investigated ILH by creating two experiments. In the first experiment, Hulstijn and Laufer's (2001) study was partially replicated in order to test the effects of different degrees of involvement load on learners' (young adults from two different proficiency levels: intensive English program and undergraduate) initial learning and retention of L2 vocabulary. In line with Hulstijn and Laufer (2001), a reading+comprehension questions task (with an involvement index of 1), a reading+comprehension questions+gap-filling task (with an involvement index of 2), and a writing task (with an involvement index of 3) were used. For the participants, who were assigned to one of the task types randomly, time allotted was the same. The findings showed a significant main effect of involvement load, as observed in higher mean scores received by the composition group in the immediate posttests. For the delayed posttests, a significant main effect for task type was found, and the mean scores for each of the three tasks significantly differed from each other, with both proficiency groups in the composition group scoring highest. The second experiment focused on learners' (young adults from two different proficiency levels: intensive English program and undergraduate) initial learning and retention of L2 words elicited by two different tasks with the same values of involvement load. This experiment was carried out to examine Laufer's (2005; as cited in Kim, 2008a) claim

that in involvement load terms, writing sentences and writing an essay are the same. The participants were randomly assigned to (a) the composition or (b) the sentence writing group, each of which had an involvement index of 3, and they were given around 40 minutes to complete the tasks. The results revealed neither significant main effects for task type and proficiency level nor an interaction between the two for the immediate and delayed posttests. Kim concluded that higher involvement required by her task was associated with better initial learning of words and also their retention. She also argued that there was some proof to claim that her tasks with the same involvement indexes were effective to the same degree. On initial learning and retention of words, language proficiency was not found to be a significant factor. One of the limitations of the study, however, as also noted by Kim, was the small sample size.

Zou (2017) aimed to explore how the evaluation load of cloze-exercises, sentence-writing, and composition-writing tasks compare with each other, as predicted by ILH. Intermediate level university students participated in this study, which also used think-aloud and retrospective interview techniques. The results showed that each task was significantly different from the others based on the scores elicited in the immediate as well as delayed posttests. This was interpreted as partial support for ILH because although the writing tasks (both with strong evaluation load and took around 35 minutes to complete) were found to be significantly superior to the cloze-exercises task (with moderate evaluation load and took around 30 minutes to complete), these writing tasks were significantly different from each other in spite of having the same involvement index. Zou recommended giving a strong evaluation load index (++) to sentence-writing and a very strong evaluation load index (+++) to composition-writing, considering that cloze-exercise gets a moderate evaluation load

index (+). The author referred to the differences between the tasks with respect to the information organization methods and pre-task planning required. The one-week time interval between the posttests was one of the limitations mentioned (cf. Kim, 2008a, who used a two-week interval between the posttests).

Martínez-Fernández (2008) looked into the impact of three different tasks on vocabulary learning as well as text comprehension. Degrees of awareness elicited by the tasks and characteristics of item types (i.e., concrete or abstract) were also investigated. Intermediate level students who were taking a college-level Spanish course were the participants of the study. The target items were used in the text four times, with the first encounter happening in a context without obvious clues for meaning unlike the following encounters. For the single gloss task (need +), the participants were supposed to read the text with L1 single glosses of the target words. For the fill-in task (need +, evaluation +), the target words were missing in the text, and therefore the participants were expected to fill in the gaps by using a wordlist which also included L1 translations. For the multiple-choice gloss task (need +, search +, evaluation +), the participants were required to read the text with L1 multiple-choice glosses. Finally, there was not bolding, glossing, or deletion of the target words in the control task. All of the participants were also asked to think aloud while reading. The results indicated different levels of awareness for the tasks with different involvement indexes, though not in the same direction as ILH would suggest. To give an example, the fill-in task elicited significantly higher awareness than the other three conditions, with multiple-choice task being one of them and having the highest involvement index. The tasks with different involvement indexes were found to have differing effects on vocabulary learning; however, these were not in line with ILH's predictions. Some of the significant effects observed included the

supremacy of the fill-in group over the multiple-choice gloss group (based on the scores of meaning production and recognition tests) and a lack of difference between (1) the multiple-choice gloss group and the single gloss group and also between (2) the single gloss group and all the other groups. Concerning time-related effects, it was the single gloss task which elicited a significantly higher loss from the immediate posttest to the delayed posttest (based on the scores of meaning production and recognition tests). As for item types, the findings revealed that all tasks yielded significantly better results for concrete nouns (as opposed to abstract nouns), and, for all types of tests, abstract nouns elicited significantly lower retention scores. With regard to text comprehension, Martínez-Fernández argued that the multiple-choice gloss task, in comparison to single gloss and fill-in tasks, could be detrimental to comprehension. The author concluded that this study did not support ILH. The small number of target word items (as noted by the author) as well as the sample size should be taken into consideration when interpreting the results of this study.

Bao (2015) aimed to compare the effects of word-focused output tasks on receptive and productive vocabulary knowledge with that of a control task. In addition, the possible effects of output-induced involvement loads on task effectiveness were investigated. As the reading text, this study used one sentence for each target word, after which the gloss for that target word was given. A four-point scale was used to check the participants' comprehension of each sentence. After reading the text, the first-year, university level participants were given one of the following five task types: (1) the control task (need -, search -, evaluation -), (2) the definition, (3) combining, (4) translation tasks (all with need +, search -, evaluation +), or (5) the writing task (need +, search -, evaluation ++). Time on task was not

held constant; however, attention was paid not to confound time on task with involvement load while explaining task type effects on learning. The results showed that, excluding the comparison between the combining and control tasks for productive knowledge of words, all the output tasks elicited significantly better results than the control task did, for both receptive and productive knowledge of the words. As for the effectiveness of different tasks, the findings revealed that the definition task was better than the other output tasks with regard to the receptive knowledge of the words. Concerning the productive knowledge of the words, the definition, translation, and writing tasks were significantly better than the combining group, although they did not differ from each other significantly. Bao stated that ILH was partially confirmed by this study. The binary categories of moderate versus strong need were also questioned by the author who argued that this "may obscure the finer differences in the strengths of need produced by different vocabulary learning tasks, or ignore the inseparability of the strength of need from the nature of a vocabulary learning task per se" (p. 92). In the context of deciding why a given task is better than another, one of the concluding remarks included the suggestion of taking into consideration "word encounter frequency, contextual clueing, and learners' awareness of the target words" (p. 93).

Yang, Shintani, Li, and Zhang's (2017) study examined how vocabulary learning as a result of post-reading word-focused tasks was affected by task-induced involvement and working memory. All of the three groups of participating university students (first-year, advanced-level English majors) were given a reading comprehension task and another task following it (completed within the same amount of time). The tasks that followed were (1) essay question task (need -, search - , evaluation -) for the comprehension group, (2) sentence writing (need +, search -,

evaluation ++), and (3) gap-filling (need +, search -, evaluation +). There was also a control group which took the pre- and posttests only. The results showed significant main effects for time and group and also a significant time and group interaction. Based on the results of the immediate posttest, which fully supported ILH, the sentence writing task elicited better performance than the remaining tasks did; the gap-filling task was superior to the comprehension and control tasks. As for the delayed posttest, the sentence writing task elicited significantly better results than the comprehension and control tasks did, though this superiority was not valid over the gap-filling task which was also found to be better than the comprehension and control tasks. The three experimental groups did not differ from each other significantly in the working memory test. Working memory was found to be a significant predictor of learning based on the immediate posttest scores for comprehension and gap-fill tasks, but not for the sentence writing task. For the delayed posttest, none of the three tasks was associated with a significant predictive model. In their discussion of the lack of a significant difference between the sentence writing and gap-filling groups in the delayed posttest, Yang et al. mentioned the possible influence of the extent of the level of processing which occurred under these task conditions requiring evaluation. With regard to the larger effect size reported for the gap-filling task (compared to sentence writing in the delayed posttest), the authors pointed out the following possibility:

> We speculate that this is because in the Sentence writing group, the meaning of the words is provided and the participants need to access the form-meaning mapping only once when they write a new sentence with the target words. However, when completing the gap-fill activity, the participants need to refer to these words repeatedly to assess the meaning and appropriateness of the words for the context. The repeated encounter with the target words and evaluation of the context may enable participants to better memorize the meaning of the target words in the long term. (p. 46)

Yang et al. (2017) concluded that ILH was partially supported by their study.

Eckerth and Tavakoli (2012) aimed to examine how frequency of word exposure and elaboration of word processing affected the learning and retention of L2 vocabulary. The participants were the students who were taking an advanced level pre-university English for academic purposes course. Three different reading tasks used were (1) reading with marginal glosses (low involvement index: need +, search -, evaluation -), (2) gap-filling by using a word list (medium involvement index: need +, search -, evaluation +), and (3) reading with marginal glosses and writing a composition (high involvement index: need +, search -, evaluation ++). Each reading text had ten target words, with five of them appearing just once, and the others five times each. In the three-week intervention period, the participants read the same reading text with the same target words but under a different task condition (i.e., 1, 2, or 3) each week. The immediate posttests given after each task measured the knowledge of the ten target words covered in that task session, whereas the delayed posttest, which was given three weeks later than the last immediate posttest, included all of the 30 target words (10x3). The findings revealed, for initial word learning, significant main effects for involvement load and exposure frequency and also a significant interaction between them. Learning gains benefitted from increasing encounters and involvement load. With regard to the types of word knowledge investigated (i.e., active recall, active recognition, passive recall, and

passive recognition), no significant interactions were found for exposure frequency or involvement load. As for word retention, one significant effect for involvement load was found. Although Tasks 1 and 2 were not significantly different from each other, Task 3 was significantly different from them. Exposure frequency was not found to have a significant effect. Although there was no significant interaction between word knowledge type and exposure frequency for the retention of the target words, a significant interaction was found for the type of word knowledge and involvement load. Additional analyses for Task 3 showed that the test of active recognition was significantly different from the remaining three. Finally, when it comes to the effects of time, a significant interaction between word knowledge type and time was found. Word recognition (active recognition more specifically) was found to have higher retention scores when compared to other knowledge types. The losses in gain scores between the posttests were equally significant across all of the knowledge types. The authors argued that the effects of involvement load were more durable, in time, when compared to those of frequency. Eckerth and Tavakoli, who concluded that ILH was strongly supported by their study, commented that "the main educational appeal of the Involvement Load Hypothesis is its potential instructional applicability: a formula for teachers to use to better manipulate and foster their students' vocabulary learning" (p. 244). They further questioned the construct validity of ILH at the time. Two of the issues they raised included a lack of control over whether participants used the glosses and previous studies' failure to express the intensity of the challenge required to choose the correct word meaning for fill-in reading tasks.

Nassaji and Hu's (2012) study examined the relationship between task-induced involvement and L2 lexical inferencing. University level L2 learners in

Canada, with two to three years of stay, were the participants of the study. Three versions of a text were created: (1) the text with multiple-choice glosses (low involvement load: need +, search -, evaluation +) which required the participants to infer the meanings of the target words by choosing one of the options given, (2) the normal text (moderate involvement load: need +, search +, evaluation +) which required inference without any options given, and (3) the text with derivationally different target words (high involvement load: need +, search ++, evaluation ++) which required inference as well as modifications to word form. Among the findings was a significant interaction effect between text type and strategy type, which meant that the number of different strategy types used by the participants changed with text type. The tendency of the participants working on the high involvement load text was to use word-based strategies more, whereas the tendency was more towards context-based strategies for the low involvement load text. Similar mean scores were elicited for correct inferences across the text types. As for the retention of the successfully inferred target words, a significant main effect of involvement load was found. The three texts significantly differed from each other in terms of the retention scores elicited (i.e., high > low > moderate), with the primary difference being between the moderate and high involvement load texts.

Pichette, de Serres, and Lafontaine (2012) compared reading and writing of sentences to see their effects on the incidental learning of L2 vocabulary. Word concreteness was also integrated into the design as an effort to contribute to ILH. Intermediate and advanced level university students, for the writing task, were expected to write three sentences, each of which had to include the target word. The reading task included three sentences in which the target word appeared as a subject, a direct object, and an indirect object. Time on task was not limited; the reading task

took around half of the time taken for the writing task. Four task conditions (i.e., writing concrete words, writing abstract words, reading concrete words, and reading abstract words) were used in this within-subjects design. After task completion, the participants were, immediately and one week later, asked to write the target words for the L1 definitions provided. The results showed that vocabulary gains, for immediate and delayed recall tests, were significantly lower in the reading task. Vocabulary gains, as a result of reading as well as writing, were significantly lower in the delayed recall test. As for word concreteness, a significant difference between abstract and concrete words was found only in the immediate recall test, with abstract words eliciting lower scores. Scores for all target words (i.e., abstract and concrete) were significantly lower in the delayed recall test. Finally, it was further reported that by the time of the delayed recall test, the supremacy of the writing task vanished for the abstract words. The authors commented that the results were in line with the predictions of ILH, as "generally, writing a text may lead to significantly higher recall than reading if enough time is allocated for each task, writing being intrinsically longer than reading for the same amount of language" (p. 77).

To sum up, some of the above studies differed in the degree of evaluation their tasks induced (Hulstijn & Laufer, 2001; Keating, 2008) and others showed qualitative as well as quantitative differences across the tasks they investigated by increasing item numbers within a single task (Folse, 2006; Jahangiri & Abilipour, 2014). Some evidence has been found for ILH (Eckerth & Tavakoli, 2012; Huang, et al., 2012; Hulstijn & Laufer, 2001; Karalık & Merç, 2016; Keating, 2008; Kim, 2008a; Yanagisawa & Webb, 2021; Zou, 2017). However, when time on task was controlled, a step that Hulstijn and Laufer (2001) did not intentionally take, differential effects of task-induced involvement seemed to disappear (Jahangiri &

Abilipour, 2014; Folse, 2006; Keating, 2008). According to Keating (2008), research set out to test the claims of ILH focused only on advanced learners, tested the passive knowledge of words, and did not take time on task into consideration in the analyses of task performance, which together constitute the limitations of these studies. Furthermore, Hu and Nassaji (2016) concluded that ILH worked less effectively than TFA. Schmitt (2008) pointed to the strength of ILH from a materials design perspective, and its weakness form a learner perspective.

All in all, it seems that task type as well as time taken to complete tasks constitute important parameters in predicting the retention of L2 vocabulary.

2.2 Vocabulary and reading

Paribakht and Wesche (1997) compared the two conditions which required the participants to read their assigned readings and answer comprehension questions but differed from each other in terms of the final activity. While in one of the conditions the participants completed vocabulary exercises on their assigned readings, in the other condition, the same participants read additional texts including the target words from their assigned readings and completed comprehension exercises. Time spent on these activities was the same for both of the conditions and the Vocabulary Knowledge Scale (VKS) was used to track the participants' gains in their knowledge of the target words. Although significant increases were observed for all target word categories as a result of completing both of the tasks, reading plus vocabulary exercise condition led to significantly higher gains than the reading condition. In terms of "quantitative (reflected in the number of words known to some degree versus not known) and qualitative (increased 'depth' of knowledge of given words)" (p. 189) gains, more gains were observed for the reading plus vocabulary exercise

group. Paribakht and Wesche referred to “the amount and variety of mental processing required” (p. 196) while discussing the potential causes behind the effectiveness of the reading plus vocabulary exercise condition. One noteworthy feature of this study was the categorization of the vocabulary exercises. The authors compiled a set of exercise types from textbooks on L2 vocabulary teaching and classified them in an assumed ranking of mental processing required: *selective attention*, *recognition*, *manipulation*, *interpretation*, and *production* (Paribakht & Wesche, 1996). Replacements made with the target words, a technique used in the current study, were used in this study.

Laufer (2003) criticized the proposition that reading is the main contributor of L2 vocabulary acquisition by referring to four postulates this argument is built upon. She argued that learners do not always notice the words unknown to them in a text and even when they do so, guessing the correct meaning may not be possible (see also Laufer, 1997). Laufer further suggested that retention of the words guessed from context is not guaranteed and the amount of reading that is required for retention of new words may be too excessive to undertake in the language classroom. Based on the results of three experiments, Laufer showed that engaging with words in productive word-focused tasks produced better results for recalling these words than being exposed to them as a result of reading a text.

Hill and Laufer (2003) used three tasks which differed from each other in terms of how the target words embedded in a short text would be engaged by their young adult learners of English with L1 Mandarin or Cantonese. The first task (message-oriented task) measured the participants’ comprehension of the test with yes/no questions involving each of the target words. For the second task (form-oriented comprehension task) participants saw the target word on a screen and were

asked to choose, among four options provided to them, the meaning of this word. As for the third task (form-oriented production task) participants were this time given a synonym or paraphrase of the target word and were required to choose the correct word from among options. Hill and Laufer used a computer program which timed participants' task completion and also kept track of what types of information they used for the target words (i.e., meaning of the word in English, Chinese translation, extra dictionary information concerning part of speech, preposition following the verb, and an example sentence, pronunciation of the word) and how many times they used these. The message-oriented task elicited significantly lower vocabulary scores than both of the form-oriented tasks in the immediate recall test, but it produced significantly lower scores than the form-oriented production task only in the delayed recall test. Time taken to complete the tasks did not differ across the tasks significantly. The message-oriented task, again, differed from both of the form-oriented tasks, with significantly lower number of dictionary activity. While participants completing the form-oriented production task used the translation function more, participants assigned to the message-oriented task and form-oriented comprehension task referred to the English meanings of the words. Hill and Laufer concluded that for measuring the effectiveness of a given vocabulary learning task, a critical criteria is "the amount of word-related activity that the task induces" (p. 104).

Laufer (1989) investigated the relationship between the amount of words learners could understood in a text and text comprehension, with scores above 55% considered acceptable reading comprehension. Results showed that learners with lexical coverage scores (i.e., percentage of the word tokens known to a learner) of 95% minimum had significantly higher reading comprehension scores than those who had lexical coverage scores of 94% and lower. Laufer (1989) proposed this,

95%, as a lexical threshold of reading comprehension, arguing that "lower lexical coverage will be associated with unsatisfactory more often than with satisfactory comprehension" (pp. 319-320). Hu and Nation (2000), however, found that a larger lexical coverage of 98% could be required for reading comprehension. Hu and Nation further noted that although some participants achieved reasonable comprehension with 90% and 95% lexical coverage, they were not among the majority.

In another study focusing on learners' reading comprehension again, Laufer (1992) found a significant relationship between her participants' reading comprehension scores and their vocabulary sizes. It was reported that the knowledge of 3,000 word families was the threshold for reading comprehension. Although differences in reading scores were observed between 3,000, 4,000, and 5,000 vocabulary levels, it was the difference between the 2,000 and 3,000 levels that was statistically significant.

Nation (2014) found that it would be necessary for learners to read just a little more than 300,000 words to encounter most of the words in the 3rd 1,000 word families 12.6 times on average. For the 9th 1,000 word families, this number increased to 2,956,908 running words. Nation further calculated the amount of time it takes to read texts of such numbers. Hypothesizing that a learner reads 200 words a minute per school week (40 weeks a year), he found that it would take 38 minutes of reading per week for the 3rd 1,000 word families and 6 hours and 10 minutes for the 9th 1,000 word families. Nation argued that these figures were "manageable amounts of reading in terms of the time needed" (p. 8); however, he also warned that these calculations assumed that the texts were suitable for learners' level, with the necessary percentage of known vocabulary occurring in the texts. Nation emphasized

the difficulty posed by most of the texts in graded readers beyond the 3000 word level. He commented that, for learners with vocabulary sizes lower than 9,000 word families, "[u]nsimplified text clearly provides poor conditions for reading and incidental vocabulary learning" (p. 9). Overall, it was reported that learning vocabulary through extensive reading is possible with suitable materials.

Grabe's (2009) predictions for the amount of words that can be learned through instruction seem to be based on the amount of instruction:

> Unlike in more common settings of 3-6 hours of foreign-language instruction per week, it is possible to argue that, in fairly intensive instruction (12-20 hours per week), 2,000 words per year (50 words per week x 40 weeks) could be taught directly to L2 learners. (p. 281)

Grabe further pointed to the importance of word-learning strategies for learners who have mastered the most frequent 2,000 words, and the necessity of an improved ability to learn words on the part of these learners whose previous word knowledge can facilitate the acquisition of new words.

Laufer and Rozovski-Roitblat (2015) examined how the number of encounters and activity type engaged interacted in the learning of new vocabulary incidentally. Three different tasks (10 different target words in each; given under three 'number of encounters' conditions) were used in the study: Reading only (R; 6&9 or 12&15 or 18&21 encounters), Reading plus focus of form (F; 2-3 or 4-5 or 6-7 encounters), and Reading with one encounter in the text plus focus on forms (1 + Fs; 2-3 or 4-5 or 6-7 encounters). The R task included the target words but did not provide extra information about these words (i.e., the participants would need to infer word meanings from the context). The participants were required to write a short summary for the texts. As for the F task, the participants saw the target words in the texts, and they could use a dictionary. They were also given 10 comprehension

questions (i.e., multiple-choice or true-or-false). Finally for the 1+ Fs task, the participants saw the target words once in each text, and they engaged in post-reading vocabulary focused exercises through which they worked with the target words multiple times. The numbers of the exercises were not the same across the tasks with different numbers of encounters. In the R condition, two extra texts that included the target words were also read in each session. This meant that this task tripled the number of encounters which were possible in the F and 1+Fs tasks. Overall, the R group read 21 texts in comparison to F and 1+Fs groups, which read 7 texts each. Intermediate level high school graduates enrolled in pre-academic English courses were randomly assigned to three task conditions. The posttests, which were given two weeks after the end of the treatment, targeted word recall and recognition (active as well as passive) knowledge. The results showed significant task type effects on each type of word knowledge tested in every 'number of encounters' condition. 1+Fs was found to be significantly superior to F and R for all types of word knowledge in each 'number of encounters' condition. In terms of active recall and recognition, similar results were found between F 2-3, 4-5, and 6-7 and R 6&9, 12&15, and 18&21 respectively. With regard to passive recall, F 2-3 elicited significantly higher scores than R 6&9. As for passive recognition, F 2-3 and 6-7 were significantly better than R 6&9 and 18&21 respectively. Analyses focusing on the "number of encounters" conditions revealed that increasing encounters caused a significant improvement in retention scores for all types of word knowledge for 1+Fs, three of the types for F (excluding active recall), and again three for R (excluding passive recognition). In the 1+Fs condition, contrary to R and F, every increase in encounters yielded significantly better performance (with the exception of the increase from 2-3 to 4-5 for the passive recognition of words). Of the 72 possible combinations of task

type and number of encounters, one striking finding showed that the lowest number of encounters in 1+Fs did not significantly differ from the highest number of encounters in F or R for recall knowledge (active and passive); however, for recognition (active and passive) it was significantly better. Finally, in terms of relative contribution to learning, task type was found to be superior than number of encounters. Laufer and Rozovski-Roitblat commented that, according to the results of their study, "what learners do with the word may be more important than how many times they come across it, since it is the nature of the task that determines how effective multiple encounters will be" (p. 707).

2.3 Reformulation, revision, and editing

Grabe (2009), in a discussion of standardized reading assessment, referred to the sensibility of the idea that "reading-assessment tasks and task types should change with growing L2 proficiency" (p. 357). Grabe further noted that novel methods to assess the knowledge of words, in the area of L2 reading assessment, have not been investigated in detail.

Studies focusing on target word retention from a task-induced involvement perspective have focused on groups of tasks such as [reading comprehension, comprehension plus fill-in, composition writing] (Hulstijn & Laufer, 2001), [reading comprehension, comprehension plus fill-in, sentence writing] (Keating, 2008), [reading with multiple-choice, reading and definition choice, reading plus fill-in, reading and rewording sentences] (Hu & Nassaji, 2016), [one fill-in-the-blank, three fill-in-the-blanks, sentence writing] (Folse, 2006), and [fill-in with glossary, fill-in by searching, retelling with glossary, and retelling by searching] (Karalık & Merç, 2016). This study, however, used a text reformulation task. Such a choice was aimed

towards consciousness-raising purposes, from which L2 productive skills, especially L2 writing production, would benefit considerably. Moreover, the mere occurrence of low-frequency words in a style section article, as was used in this study, could be convincing evidence for learners that such words are worth learning or are even essential for their vocabularies. Thornbury (1997) noted that reformulation activities foster conscious-raising at various levels including lexis and discourse among others and stated the following:

> rather than simply correcting a student's composition, which usually involves attention to surface features of the text only, the teacher reformulates it, using the content the student has provided, but recasting it so that the rewritten draft approximates as closely as possible to *a putative target language model* (emphasis added). (p. 327)

Although similar in the conceptualization behind it, the way texts were reformulated in this study deviated from the above description of reformulation activities in some aspects. First, the source texts were not written by the participants themselves. This was because a common base text was required to ensure the relevance of the predefined target words for the reformulation task. Second, those who reformulated the texts were not teachers but the participants themselves. The text (i.e., a style section article) higher in lexical richness measures worked as an embodiment of the putative target language model quoted above, and all the activities during task completion were either directed towards it or departed from it. Still, reading skills were involved throughout the whole process as participants needed at least a sentence-level understanding of their assigned text in order to make judgments about which words (i.e., those presented in the wordlist) would replace which ones (i.e., those embedded in the texts given).

Swain and Lapkin (2002) showed how reformulation could serve as "an effective technique for stimulating noticing and reflection on language" (p. 298)

when made on a piece of text written by learners who engaged in collaborative dialogue, compared it with their original version, and reflected on their language choice.

Polio, Fleck and Leder (1998) distinguished between revision and editing, explaining that the former encompasses all types of changes which may relate to the linguistic, organizational, content-related make-up of a previously written text whereas the latter, as an element of revision, particularly concerns the changes made at the level of a sentence.

2.4 Incidental/Intentional learning and focus on forms

Hulstijn's (2003) extensive account of incidental and intentional learning incorporates the uses of the terms in the psychological literature and the L2 learning literature. Within the former field, "incidental and intentional learning refer, strictly speaking, only to the absence or presence of an announcement to participants in a psychological experiment as to whether they will be tested after the experimental task" (Hulstijn, 2003, p. 356). While such a definition might naturally raise questions in one's mind, the following explanation of the way this definition is used seems to clear the confusion:

> … incidental learning has acquired the status of a tool in the cognitive psychologist's experimental research kit to investigate some way or ways of information processing as intended by the investigator, not contaminated by ways of information processing not intended by the investigator. The presence or absence of an intention to learn does not figure as a theoretical construct in any current theory of human cognition. (Hulstijn, 2003, p. 356)

From an L2 learning perspective, according to Hulstijn, incidental and intentional learning might be differentiated in terms of how attention is allocated: "not deliberately geared toward an articulated learning goal" in the former and "deliberately directed to committing new information to memory" in the latter (p.

361). To reiterate, now from a directly vocabulary-oriented perspective, a similar argument remains valid. According to Laufer (2003), "[i]ncidental learning does not mean that the learners do not attend to the words during the task. They may attend to the words (for example, using them in sentences, or looking them up in the dictionary), but they do not deliberately try to commit these words to memory." (p. 574).

When it comes to intentional and incidental learning, what is intentional for one student might not be the same for another (Yanagisawa & Webb, 2021). Similarly, "different learners are likely to have different intentions when engaging with input which may include obtaining information, understanding the message, learning language features (grammar and vocabulary), or simply finishing (or appearing to finish) an assigned task" (Webb, 2020, p. 225).

As Keating (2008) pointed out, although ILH was not situated in the context of form-focused instruction in the first place, "when vocabulary tasks are contrasted in terms of the mental effort they induce, it is apparent that increased involvement load generally entails greater focus on form (where *form* refers to lexical items)" (p. 368). In line with this, this study drew on a Focus on Forms (FonFs) approach to the learning of L2 vocabulary, using tasks that incorporated what Laufer (2005) refers to as "task related FonFs activities" (p. 238).

2.5 Task interactiveness and experimental tasks

Interactiveness, a component of Bachman and Palmer's (1996) model of test usefulness, refers to "the extent and the type of involvement of the test takers' individual characteristics in accomplishing a test task" (p.25). Bachman and Palmer suggested that *language ability*, *topical knowledge*, and *affective schemata* constitute

the most pertinent individual characteristics for language testing (see Figure 1), and that test tasks which necessitate that a link be established between the test (e.g., topical content) and test taker's previous knowledge (e.g. knowledge of the topic) could be considered as more interactive in comparison to those lacking such a feature.

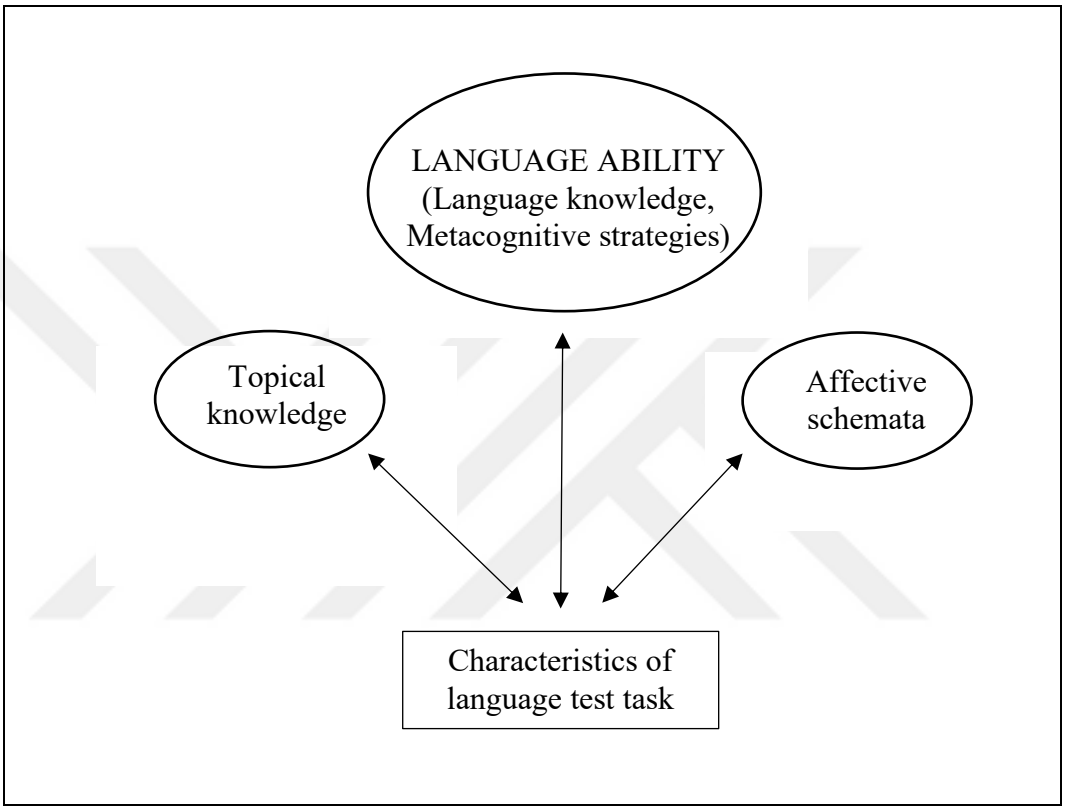

Figure 1. Interactiveness (Bachman & Palmer, 1996, p. 26)

Topical knowledge constitutes one of the components of Grabe's (2009) classification of knowledge which was motivated by a desire to demonstrate the intricacies of background knowledge. Grabe suggested with certainty that readers who possess "considerably more background knowledge on a topic read a text differently and more efficiently" (p. 74), yet he remained cautious about the effects of background knowledge on reading, concluding that "it is probably safe to say that it is not a concept that is well-specified to this point in reading research" (p. 76).

Sanz (1997) pointed out that creating a variety of assessment tasks is necessary in order to better identify the outcomes of instruction and that incorporating mode and amount of production in such tasks could produce insightful interpretations as to differences found in learners' production in line with the demands of these tasks.

2.6 Word frequency and word families

Frequency, as Vilkaitė-Lozdienė and Schmitt (2020) argued, "is a good guiding criterion for word selection as it is very straightforward and objective" (p. 82). Similarly, the reason for categorizing words as high- and low-frequency is to allow teachers to decide on the words which deserve teaching time in the classroom and those which do not (Nation, 2001b). Still, it seems important to approach word frequency information cautiously. For example, Nation (2007) suggested that the frequency lists based on the British National Corpus (BNC) display the "truism in corpus linguistics – the composition of the corpus determines the nature of what is drawn from it" (p. 38) and he further argued that these lists represent a compilation of formal British English as used by adults. In learning of words, nevertheless, the significance of frequency "is as near to a fact as it is possible to get in L2 acquisition" (Milton, 2009, p. 242). Raising also the problems involved in a frequency-informed approach for lexical richness (i.e., for gauging learners' vocabulary level) Bardel (2016) noted that frequency seems to be one of the most useful measures. In a study they conducted, Arnaud and Savignon (1997), for example, used two different frequency lists so as not to rely solely on British or American English.

Another issue that pertains to the discussion of frequency relates to their categorization. As Kremmel (2016) stated, all types of categorizations will likely be a matter of choice and their "boundaries will unavoidably create anomalies" (p. 980). The boundaries of what is high-frequency is still an issue open to debate, and a similar lack of precision exists when it comes to what is low-frequency (Vilkaitė-Lozdienė & Schmitt, 2020). Nation (2001b) suggested that 2,000 words can be taken as high-frequency. Schmitt and Schmitt (2014), on the other hand, argued that high-frequency words should include the first 3,000 word families, and the 9,000+ word families should be considered as low-frequency vocabulary. Schmitt and Schmitt (2014) further claimed that the binary high- and low-frequency categories are hard to maintain because some word families following the 3,000 (i.e. high-frequency) word families are "clearly too useful to be written off as low-frequency vocabulary" (p. 493). They suggested that the word families between 3,000 and 9,000 word families be named *mid-frequency* vocabulary, and by referring to various studies, they also discussed the benefits of learning mid-frequency words (e.g., for reading English-language textbooks and watching movies). Kremmel (2016) supported the use of smaller categories (i.e., 500 lemmas rather than 1,000) for higher frequency words while suggesting categories larger than 1,000 lemmas for lower frequency words. Cobb and Laufer (2021) introduced the Nuclear Family List 7 (NFL7), which they argued is a more compact word list and can be used with learners from beginner to upper intermediate levels. What makes NFL7 a different word list, according to the authors, is that only the most frequent word family members (which can be inflected as well as derived forms) realizing a minimum of 7% of the occurrences of the entire family were allowed in the lists.

Bauer and Nation (1993) explained word families as follows:

> From the point of view of reading, a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately . . . . The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort. (p. 253)

Bauer and Nation maintained that teaching and learning is greatly affected by the selection of the words which are allowed in a given word family.

Highly related to the discussion of word families, as part of the bigger discussion of word frequency, is other possible categorizations of words (see Cobb & Laufer, 2021; Vilkaitė-Lozdienė & Schmitt, 2020 for a discussion of word families and lemmas.). According to Kremmel (2016), for vocabulary tests, it appears that counting units based on word families are not good alternatives, but lemmas could work better with more certainty that a particular word which appears in the test actually signals that the remaining members in its group are known. For Nation (2014), as long as receptive knowledge is concerned, the most appropriate unit of analysis is word families.

CHAPTER 3

METHODOLOGY

This study aimed to explore the effects of task type on the retention of low-frequency L2 vocabulary. With the two tasks designed for this study, which differed only slightly from each other and thus had different involvement indexes, it was hoped that an answer would be found to Laufer and Hulstijn's call for extending the scope and depth of their hypothesis by suggesting, respectively, a fourth component or "more precise definitions of the involvement components" (pp. 22-23). To this end, this study adopted a lexical richness (Read, 2000) perspective in its approach to task-induced involvement and aimed to investigate the effectiveness of two tasks, which varied in the *evaluation* they induced, on advanced learners' retention of low-frequency words. This study sought answers to the following research questions:

i. Do advanced learners achieve better retention (early and long-term) of low-frequency words after completing a more involving reading+replacing task?

ii. Are low-frequency words encountered in a reading+replacing task actively recalled in a sentence revision task?

iii. How do the participants perform in relation to task and individual word features before, during, and after completing the reading+replacing task?

For the first research question, it was predicted that the more involving reading+replacing task would lead to better retention of the target words, both in the short and long term (Hypothesis 1). For the second research question, no predictions concerning the extent of correct target word form use were made (Hypothesis 2).

Similarly, for the third research question, no predictions were made; however, task design and methodological factors were expected to show their effects on the results (Hypothesis 3).

In the analyses of the data elicited from the participants which will be presented in this chapter, the existence of missing data for participant background information or test items should be kept in mind.

3.1 Participants

The participants were 53 (Group A = 26 and Group B = 27) undergraduates at a Turkish state university where the medium of instruction was English. Five of the participants were absent in the second sessions, which reduced the number of participants for whom the delayed posttest results available to 48 (Group A = 23 and Group B = 25). 15 male and 38 female participants in the present study were predominantly 20- and 21-year-olds (M = 21.39, SD = 2.244), English language education majors, and in their second or third year of study at their degree programs, which are shown in more detail in Figure 2.

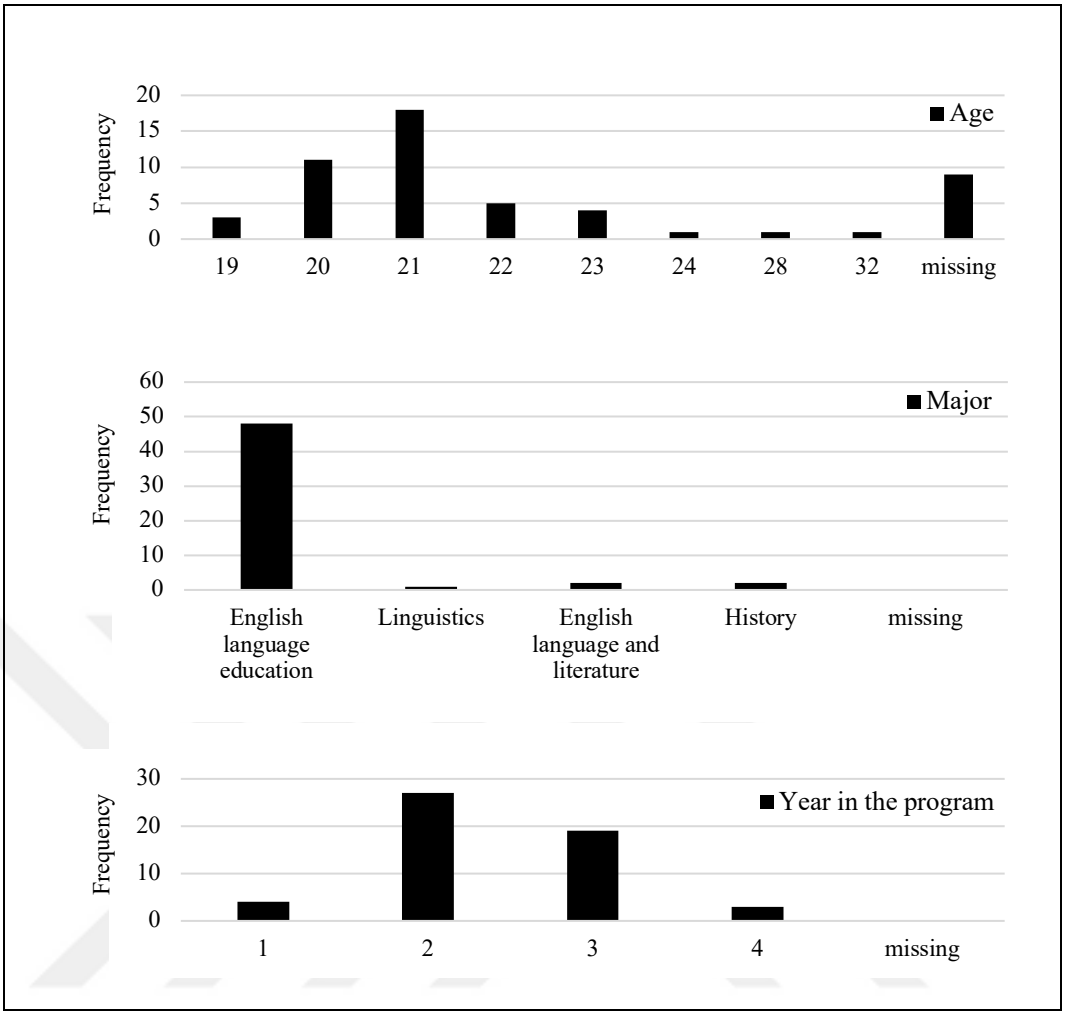

Figure 2. Participant age (top), major (middle), and year in the degree program (bottom)

Among all the participants, 58.5% stated that they had received preparatory year English language education at university. The number of semesters spent at the preparatory program was two for the majority, i.e. 49.1%. The majority of those, 84.4%, who reported studying in the preparatory English program completed it with As and Bs (see Figure 3).

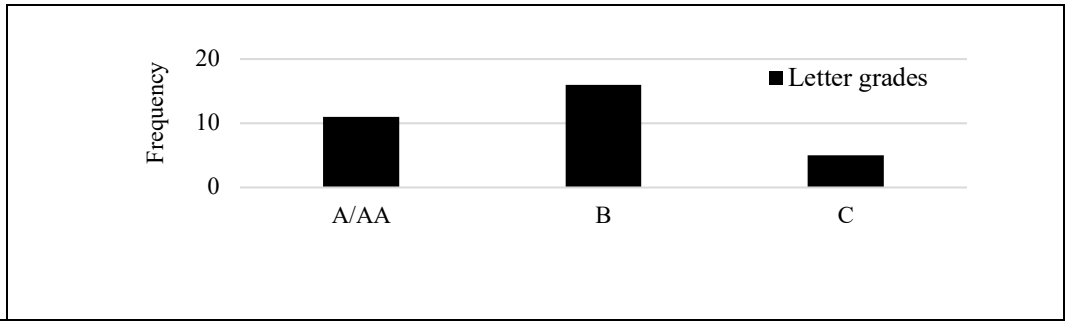

Figure 3. Pass scores for those completing the preparatory English program

Those who stated that they had not studied in the preparatory English program, again, reported having A- and B-level grades mostly as exam results, i.e., 89.5% (see Figure 4).

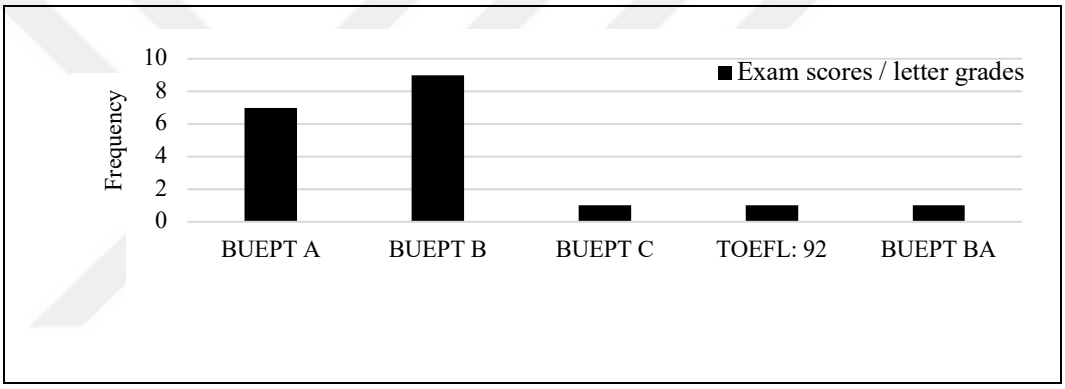

Figure 4. Exam results received by those who did not attend the preparatory

(a) TOEFL® IBT or TOEFL® IBT (Special Home Edition Test) score of 79 along with a Writing section score of 22 and (b) IELTS Academic (Paper-based/ Computer-delivered) score of 6.5 along with a Writing section score of 6.5 are the minimum pass scores equivalent to that of BUEPT, the in-house proficiency test given at the institution ("Eşdeğer Sınavlar" [*Tests Equivalent*], n.d.). For those who have received the general score required in these tests but failed in the writing section, there is the option of passing the in-house writing test and proving their proficiency. As such, in TOEFL® IBT terms, the participants were at least high-intermediate in writing (with the minimum pass score of 22 being two points below the advanced level) and the minimum total pass score of 79 could be met with the

minimum advanced level scores in the reading (i.e., 24), listening (i.e., 22), and speaking (i.e., 25) sections (Understanding Your TOEFL, n.d.).

As for their engagement with the English language, the majority of the participants (54.7%) chose 11-15 years as the time period they had been learning English for or as the number of years passed since they had started learning English (see Figure 5).

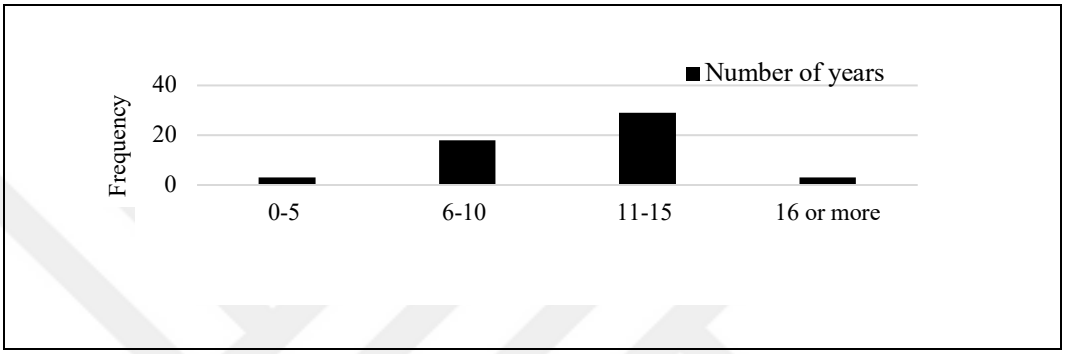

Figure 5. English language learning history

Only four of the participants (7.5%) stated that they had stayed in a country where English is spoken as the mother tongue for more than a month.

3.2 Instruments

Some instruments used in this study underwent certain changes, sometimes more than once, to be adapted to the conditions brought about by the pandemic. Unfortunately, these changes sometimes necessitated the removal of a feature before the data collection started, and at other times, changes were made even after data collection started. One of the biggest changes of this kind was the delivery method.

3.2.1 A brief overview of the tasks

The materials used in this study were given over two sessions and in three parts. Table 2 displays the materials used in each session in the order they were given:

Table 2. Materials for Tasks A and B

Session and Part	Content	Forms
Session 1- Part 1	1. An informed consent form 2. A demographics form 3. A reading text and a wordlist 4. A mini-dictionary 5. A word replacement task 6. Reading comprehension questions 7. A survey of Task Interactiveness	A & B
Session 1- Part 2	8. An immediate posttest (A test of preknowledge and the VKS[a])	Single
Session 2 - Part 3	9. The Vocabulary Size Test[b] 10. Delayed posttests (A test of form recall and the VKS)	Single

[a]the Vocabulary Knowledge Scale (Paribakht & Wesche, 1997). [b]the Vocabulary Size Test (Nation & Beglar, 2007).

3.2.2 The reading texts and word lists

In order to compare the effects of tasks with different involvement load indexes on the retention of low-frequency words, two reading texts were used as part of larger tasks (i.e., Task A and B) which required the participants to make replacements with these words. As the reading text, Task A (see Appendix A) had the LowText while Task B (see Appendix B) had the HighText. These texts were provided with a word list which included 13 words, 10 necessary for successful task completion and the remaining three given as extras. The only difference between the HighText and the LowText was 10 words. While the HighText included the 10 target words which were beyond the first 2,000 word families, the LowText included the words that have a similar meaning to the target words but belong to the first 2,000 word families (with the exception of two words which were in the first 3,000 word family but exist in Turkish). The labels “High” and “Low” were given to reflect this difference between the texts: the text with a higher percentage of words beyond the first 2,000 word families was labelled the HighText and the text with a lower percentage of

words beyond the first 2,000 word families the LowText. As can be understood, these two 10-word sets were at the core of the texts. When, for example, the set of target words was embedded in the reading text, the other set (i.e., the counterparts with a similar meaning) was in the accompanying word list or vice versa. When embedded in the text, the target words (as in HighText) and their more frequent counterparts (as in LowText), were in bold type and highlighted. Although each text required a different word list for the participants to work with during task completion, the target words remained the same across the texts.

The participants working on the LowText were expected to look up the low-frequency words provided to them in the wordlist and decide on which words in bold in the text they would replace. The reverse was expected for the HighText. That is, the participants were expected to look up the low-frequency words in bold in the text and decide on which words in the wordlist they would replace. Based on Laufer and Hulstijn's (2001) discussion of *involvement*, and as displayed in Table 3, both tasks were identical in terms of the *need* (moderate) they induced because they required the participants to know the meaning of the target words and in terms of *search* (absent) as these words were glossed in a separate list (i.e., mini-dictionary) right below the reading text. The tasks differed in the *evaluation* they induced: The LowText task required the evaluation (moderate) of the target words. As such, this task was made up of what Yanagisawa and Webb (2021) reported as the most frequently examined combination of ILH components in their meta-analysis. The HighText task did not require any evaluation (absent) on the part of the participants because the target words were not supposed to be used actively in a context; rather, they were supposed to be looked up, only to be replaced with their high-frequency counterparts provided in the word list.

Table 3. Involvement Loads of the Reading+Replacing Tasks

	Need	Search	Evaluation
LowText	moderate (+)	absent (-)	moderate (+)
HighText	moderate (+)	absent (-)	absent (-)

Overall, when assigned one of the texts and asked to make the replacements mentioned above, the participants were ultimately expected to change the lexical profile of their assigned text by either increasing or decreasing the number of less frequent words in the text. In other words, a participant who was assigned the HighText, provided that she completed the task successfully, was supposed to change it to the LowText or vice versa.

The texts were based on Widianto's (2019) article entitled "Does Influencer Grammar Matter?", which was published online in the style section of the New York Times. The LowText and the HighText were created with an intention to keep the number of changes made to the original text as low as possible. All of the target words used in the tasks appeared in the original text, and the LowText and the HighText differed from each other only in the exclusion or inclusion of these target words. Words occurring in the original text with inflections were used without inflections in the texts (i.e., bungling –> bungle, dissections –> dissection, detractors –> detractor, desecrating –> desecrate, ramifications –> ramification) so as to rule out the use of affixes as clues for task completion.

Target words, their counterparts, and extra words were at the core of task design. To determine the word families of the words provided as candidates for replacement in each text in this study, as well as the extra words, Nation's (2012) BNC/COCA word family lists (25,000 words; Version 1.0.0) were taken as reference. This was done manually. Table 4 illustrates the location and word family

information for the target words, their counterparts, and extra words for the HighText and the LowText.

Table 4. Target Words, Their Counterparts, and Extra Words

	LowText				HighText			
	Text	Word Family	Wordlist (1-10: Target words; 11-13: Extras)	Word Family	Text (Target words)	Word Family	Wordlist (11-13: Extras)	Word Family
1	unclear	1k	muddled	7k	muddled	7k	unclear	1k
2	ruin	2k	bungle	10k	bungle	10k	ruin	2k
3	foolish	2k	crass	10k	crass	10k	foolish	2k
4	analysis	3k	dissection	7k	dissection	7k	analysis	3k
5	aggressive	3k	confrontational	3k	confrontational	3k	aggressive	3k
6	hater	1k	detractor	7k	detractor	7k	hater	1k
7	complain	2k	lament	5k	lament	5k	complain	2k
8	argument	2k	premise	3k	premise	3k	argument	2k
9	pollute	2k	desecrate	10k	desecrate	10k	pollute	2k
10	effect	2k	ramification	8k	ramification	8k	effect	2k
11			inert	8k			inactive	2k
12			instigate	7k			begin	1k
13			repository	7k			container	2k
		1-3k[a]		3-10k[a]		3-10k[a]		1-3k[a]

[a]Word family range.

Distributed as three adjectives, three verbs, and four nouns; the target words were *muddled* (adj), *bungle* (v), *crass* (adj), *dissection* (n), *confrontational* (adj), *detractor* (n), *lament* (v), *premise* (n), *desecrate* (v), and *ramification* (n). These words came from the third, fifth, seventh, eighth, and 10th 1,000 word families. All of the counterparts of the target words, except for *analysis* and *aggressive*, belong to the first 2,000 word families. However, having *analysis* and *aggressive* (as words from the 3,000 word family) among the more frequent words was not regarded as a problem. It was assumed that the participants knew these words as Turkish has these words, *analiz* and *agresif*, respectively, as words of French origin (Türkçede Batı

Kökenli Kelimeler Sözlüğü, n.d.). The counterparts set included *unclear* (adj), *ruin* (v), *foolish* (adj), *analysis* (n), *aggressive* (adj), *hater* (n), *complain* (v), *argument* (n), *pollute* (v), and *effect* (n).

The three extra words provided in each word list paralleled the characteristics of the 10 words they accompanied in terms of their parts of speech (an adjective, a verb, and a noun were sampled) and word family information (excluding the two cases of 3k words for one of the lists). Those which were provided with the target words, namely *inert* (adj), *instigate* (v), and *repository* (n), were from the word families beyond the first 2,000. The ones that were given along with the counterparts of the target words were *inactive* (adj), *begin* (v), and *container* (n), and they belonged to the first 2,000 word families. With an aim to create a parallelism across the texts, a parallelism (in terms of meaning) between the two sets of extra words was created.

3.2.3 The mini-dictionary

As the participants were provided with a mini-dictionary (of 10 target words and 3 low-frequency extra words, see Appendix C) with definitions retrieved from the Longman Dictionary of Contemporary English Online (https://www.ldoceonline.com/), further considerations were deemed necessary to have some control over the meaning search process. For instance, it was considered necessary to prevent the participants from seeing the counterpart words in the target words' dictionary definitions, which would potentially lead to an automatic replacement without focusing on the meaning of the target words. It was ensured that each target word did not include its counterpart in its definition. Another point of consideration was to ensure that all the target words had separate dictionary entries.

This was accomplished with an exception, namely the word *dissection* for which the verb form *dissect* was provided. This, however, was not considered as a possible source of difficulty on the part of the participants, advanced learners of English.

3.2.4 The word replacement task

This task required the participants to choose a synonym for each one of the ten words in bold in their assigned text from among 13 word options (see Appendix A for Task A, see Appendix B for Task B). In other words, for each word replacement, the participants were given the entire word list as the pool of options to choose from. Participants who made all the replacements between correct high- and low-frequency counterparts would get a full score of 10 (i.e., 10 replacement items x 1 point) for this task.

3.2.5 Reading comprehension questions

Five reading comprehension questions were provided in a multiple choice format, with four options to choose from for each item (see Appendix D). The purpose of using this test was to understand whether the text was suitable for the participants' proficiency level. Participants who answered all the reading comprehension questions would get a full score of 5 (i.e., 5 reading comprehension questions x 1 point) for this task.

3.2.6 The survey of Task Interactiveness

In order to examine Task Interactiveness (Bachman & Palmer, 1996), a five-item scale was given to the participants with five points each (1 = Strongly disagree and 5 = Strongly agree, see Appendix E). Participants who showed strong agreement to all

of the statements probing task interactiveness would get a full score of 25 (i.e., 5 Task Interactiveness items x 5 points) on this survey.

3.2.7 The tests

3.2.7.1 The test of preknowledge

Despite its use in the posttest session, this test worked as a measure of the participants' previous knowledge of the target words (see Appendix F). Hulstijn and Laufer (2001) used this kind of a measure with their participants following task completion in order to elicit if they had known the target words before their engagement in the task. The participants in this study were given the target words with two options to choose from for each word: "*I knew this word [before this study].*" and "*I didn't know this word [before this study].*"

3.2.7.2 Immediate and delayed posttests

It was important to use a measure which could cater for different degrees of word knowledge, an example of which was the Vocabulary Knowledge Scale (VKS) introduced by Paribakht and Wesche (1993) and also used in Kim's (2008a, 2008b) studies in an adapted version. A later version of the VKS (Paribakht & Wesche, 1997) consists of the self-report categories shown in Table 5:

Table 5. VKS Elicitation Scale (Paribakht & Wesche, 1997, p. 180)

I	I don't remember having seen this word before.
II	I have seen this word before, but I don't know what it means.
III	I have seen this word before, and I <u>think</u> it means _____. (synonym or translation)
IV	I <u>know</u> this word. It means _____. (synonym or translation)
V	I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

The VKS (Paribakht & Wesche, 1997, p. 180) was used both as the immediate and delayed posttest (see Appendix G). For this test, each target word was given with the 5-point scale separately. The lowest and highest score for each target word were 1 and 5 respectively.

Kim (2008a), cautioning the possible drawbacks of getting total scores on the VKS, used a three-way categorization of the scores as well, namely *1*, *2*, or *3 or more*. Similarly, in this study, in addition to the individual scoring of the scale items (i.e., five possible scores which show degrees of knowledge for each target word), dichotomous scores (i.e., 0 = meaning unknown versus 1 = meaning known) were also calculated. Unlike Kim's (2008a) three-way categorization, a two-way categorization was used and VKS scores of 1 and 2 (meaning unknown) were grouped dichotomously against 3 or more (meaning known). With this additional categorization of scores, participants who scored at least 3 for each target word would get a full score of 10 (i.e., 10 target words x 1 point) in this test.

3.2.7.3 The Vocabulary Size Test (Nation & Beglar, 2007)

Participants' vocabulary size was measured with the Vocabulary Size Test (Nation & Beglar, 2007) by directing them to the website https://my.vocabularysize.com/select/ test. In this test, each 1,000 word level is tested with 10 words; therefore, each item is actually a measure of 100 word families. Scores received are multiplied by 100 to calculate the total vocabulary size. To illustrate, correct answers to 10 items in each one of the 14 word family level leads to the full score 140 (i.e., 10 items X 14 word families). The full score 140 shows that the learner knows the most frequent 14,000 word families in the English language (i.e., 140 x 100). Nation and Beglar's report of the early studies which used the test revealed an approximate vocabulary size of

5,000-6,000 word families for non-native undergraduates functioning successfully at a university where the medium of instruction was English.

3.2.7.4 The test of form recall

In this test, the aim was to measure the participants' recall of each target word with the help of a sentence which included a counterpart used in the reading+replacing tasks (see Appendix H). For the high-frequency counterpart in each sentence, the participants were expected to provide the low-frequency target word. The majority of the original sentences for this test were obtained from Lexico (www.lexico.com) where they were displayed as example sentences when the target words were searched. Changes were made to these example sentences where necessary. Participants who provided the low-frequency target word for each sentence would get a full score of 10 (i.e., 10 form recall items x 1 point) in this test.

The immediate and delayed posttest batteries differed from each other. Table 6 shows the tests included in each battery as well as the orders in which they were presented to the participants. As can be seen, attention was paid to the order of the tests by presenting the test probing the recall of word forms before the test targeted to elicit the participants' knowledge of the target word meanings.

Table 6. Orders of the Posttests

Immediate Posttest Session	Delayed Posttest Session
1. The test of preknowledge	1. The Vocabulary Size Test[b]
2. The VKS[a]	2. The test of form recall
	3. The VKS

[a]The Vocabulary Knowledge Scale (Paribakht & Wesche, 1997, p. 180). [b]The Vocabulary Size Test (Nation & Beglar, 2007).

3.3 Procedure

Data were collected during the course of three consecutive semesters. The first round (Fall: only online participation option) functioned as piloting whereas the data from the last two (Spring: online and face-to-face participation options; Fall: only face-to-face participation option) were included in the final dataset.

3.3.1 Piloting

The data from the first round of participants ($N = 7$) who participated online via Zoom and GoogleForms were considered as the pilot data and were not included in the analysis. These participants were recruited from Academic Writing courses offered in an undergraduate program in foreign language education at a state university. Of the seven sections offered, the researcher had the opportunity to access six, either face-to-face or online. In the introductory visits, the researcher gave brief information about the study and explained the other important aspects such as the requirements and the arrangements of the sessions. Those who were interested were asked to share their email addresses with the researcher so that she could contact them later for the scheduling of the sessions. The participants were later assigned to sessions scheduled via Zoom. The Zoom meeting settings enabled waiting room, and allowed participants to chat (with host and co-hosts), rename themselves, and start video. The participants were given 120 minutes to complete the first session. Using a single PowerPoint slide, the researcher gave a list of instructions at the beginning of the session. The slide was kept on screen during the course of the study, excluding some interruptions from time to time. The participants were not allowed to unmute themselves, so communication was maintained via the chat box. The time interval between the first and the second sessions ranged between 13 and 14 days.

Revisions made concerned an item in demographic information and the wording of the instructions in the immediate posttest. After this round, all instructions were changed to Turkish and time allotted was reduced from 120 minutes to 90 minutes.

3.3.2 Data collection

Data were collected in two semesters, in spring and fall. In the spring semester, participants could join the study in person or online via Zoom. Those who preferred face-to-face participation used a computer to complete their assigned tasks via GoogleForms. Those who joined online completed the same tasks via GoogleForms in a Zoom meeting. There were cases where the participants were allowed to take the delayed posttest one day early or late, change their participation method (i.e., online or in-person) in the second session, or join with their webcam off. In the fall semester, the participants joined the sessions face-to-face, but this time they completed the tasks on print copies of the GoogleForms version of the tasks.

3.4 Data analysis

Data analysis procedures followed for the scoring of participants' performance on (1) the VKS and (2) the mixed ANOVA conducted on the preknowledge and posttest knowledge group scores will be detailed in this section.

The scoring of the VKS items was done in line with Paribakht and Wesche (1997), as also displayed in Figure 6. In the cases of unusual circumstances, further decisions were made. As discussed in Paribakht and Wesche (1997), a score of 2 was given to wrong answers for the categories of III, IV, or V. When the categories of III or IV elicited a correct synonym or translation, a score of 3 was given. The

participants were awarded 4 points when they used the word with correct meaning but with wrong grammar in a sentence. Finally, a score of 5 indicated the use of the word in a sentence with correct meaning and grammar, regardless of any errors elsewhere in the sentence.

Participants' responses to the VKS, which included items that required objective as well as subjective scoring, were scored by two raters. The researcher was the first rater. Both the first and the second rater were non-native English language instructors at a Turkish university with five (the first rater) and six (the second rater) years of teaching experience as well as a BA and an MA in English language education. Percent agreement for the raters was calculated by dividing the number of agreements by the total number of scorings. Scorings of items which required no judgment from the raters were not included in the calculations. As such, 422 agreements were found in 484 scorings, which showed 87.19% agreement between the two raters' scorings. Disagreements were resolved through discussion.

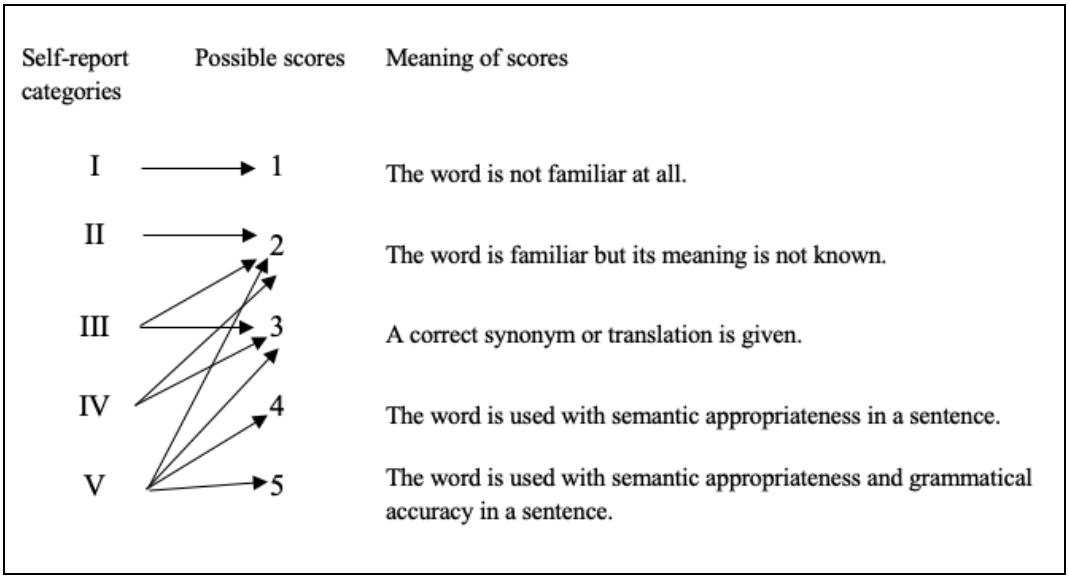

Figure 6. VKS score meanings (Paribakht & Wesche, 1997, p. 181)

The VKS scores were used in two ways: (1) original scores ranging from 1 to 5, which showed different degrees of knowledge and (2) scores transformed to

dichotomous categories of 0 and 1, meaning unknown and known respectively. Such a transformation was necessary to make the scores elicited in the posttests relatable to the ones elicited in the preknowledge test, for which the answers were in the form of binary *previously known* (i.e. a score of 1) or *previously unknown* (i.e., a score of 0).

SPSS software version 27 was used to run a 2x3 mixed ANOVA, with task type (i.e., more involving task vs. less involving task) as between subjects factor and time (i.e., previous knowledge vs. early retention vs. long-term retention) as within subjects factor. The test variables were initially checked for missing data, outliers and normality. The descriptives are given in Table 7 below.

Table 7. Descriptive Statistics for Target Word Knowledge

	N	*M*	*SD*	*Min*	*Max*
			Group A		
Test of Preknowledge	26	3.73	2.09	0	8
Immediate Posttest	26	3.35	1.90	0	7
Delayed Posttest	23	3.17	2.17	0	7
			Group B		
Test of Preknowledge	27	4.04	2.07	1	8
Immediate Posttest	27	3.78	1.99	1	8
Delayed Posttest	25	3.04	2.05	0	7

To prepare the dataset for analysis, outliers were checked. Group-by-group sums of preknowledge of the target words (time 1), immediate posttest scores for VKS (time 2, dichotomous), and delayed posttest scores for VKS (time 3, dichotomous) were calculated. Winsorizing, one of the approaches to decrease the effects of bias

mentioned in Field (2018), was used to deal with an outlier. This technique requires "replacing outliers with the next highest score that is *not* an outlier" (Field, 2018, p. 264). Skewness and kurtosis values and the Kolmogorov-Smirnov test were used to check normality. Group A's previous knowledge of the target words (i.e., Time 1), immediate posttest VKS scores (dichotomous) (i.e., Time 2), and delayed posttest VKS scores (dichotomous) (i.e., Time 3) were analyzed. Following Field (2018), in order to check "whether the [skewness and kurtosis] values are significantly different from 0 (i.e., normal) using *z*-scores" (p. 247), these scores were divided by their standard errors. *z*-scores for skewness significance tests for Time 1, Time 2, and Time 3 were respectively .103/.456 = .225, .144/.456 = .315, and .430/.481 = .893, all smaller than 1.96. This meant that skewness was not significantly different from normal. As for kurtosis *z*-scores, Time 1, Time 2, and Time 3 elicited the scores of -.922/.887 = 1.039, -.954/.887 = 1.075, and -.693/.935 = .741, respectively. Again, these showed that kurtosis was not a problem for this group. The same procedure was followed for Group B scores. *z*-scores for skewness significance tests for Time 1, Time 2, and Time 3 were respectively, .314/.448 = .700, .237/.448 = .529, and .635/.464 = 1.368. This meant that skewness was not significantly different from normal. As for kurtosis *z*-scores, Time 1, Time 2, and Time 3 elicited the scores of -.591/.872 = .677, -.827/.872 = .948, and -.195/.902 = .216, respectively. These values meant that kurtosis was not a problem for this group.

Finally, to guarantee that the two groups were not different initially, an independent samples t-test was conducted on their preknowledge of the target words. The mean difference between the two groups was not significant $t(51) = -.537$, $p = .594$. On average, Group B's reported preknowledge of the target words was higher ($M = 4.04$, $SE = .398$) than that of Group A's ($M = 3.73$, $SE = .410$).

CHAPTER 4

RESULTS

In this chapter, the answers found for the research questions will be reported.

4.1 Research question 1

The first research question aimed to examine whether the two reading+replacing tasks of different involvement load indexes used in the study led to differential gains in participants' knowledge of the target low-frequency words. A mixed ANOVA, where the different task groups (i.e., A = LowText, more involving and B = HighText, less involving) were modelled as the between subjects factor and the different times the knowledge of the target words were measured (i.e., Time 1 = preknowledge, Time 2 = knowledge at the time of the immediate posttest, and Time 3 = knowledge at the time of the delayed posttest) were the within-subjects factor, was run to find any main and interaction effects. The results showed that there was no significant main effect of group, $F(1, 46) = 0.011$, $p = 0.916$. This meant that when the test scores elicited in different times are ignored, the more involving and the less involving tasks led to similar vocabulary gains. This finding did not support Hypothesis 1 which associated better retention with the more involving task. There was a significant main effect of time, $F(2, 92) = 3.701$, $p < 0.05$, which meant that when the remaining variables were ignored, knowledge of the target words differed. No significant effects were found for time x group interaction.

With the significant main effect of time elicited, the contrasts were further visited. It was seen that the participants' preknowledge of the target words was significantly different from their knowledge at the end of the study, with the former

condition eliciting higher scores than the latter, $F(1, 46) = 7.858$, $p < 0.05$, $r = 0.3819$[3]. In contrast, no significant difference was found between the participants' knowledge of the target words when tested immediately after doing the tasks and around two weeks later, $F(1, 46) = 3.797$, $p > 0.05$.

4.2 Research question 2

The second research question asked whether the low-frequency words encountered in a reading+replacing task could be recalled in a sentence revision task. To answer this question, for which no hypotheses were made, the number of participants, who provided the correct low-frequency target word for each high-frequency counterpart, was calculated (see Table 8).

Table 8. Number of Participants Recalling the Target Words

	effect	unclear	aggressive	complain	argument	foolish	ruin	pollute	analysis	hater
Group A	0	2	2	2	4	1	0	1	4	4
Group B	2	1	1	2	3	1	1	0	2	5

Overall, the participants did not tend to use the low-frequency target words in appropriate contexts provided to them. The target word which was recalled the most was *detractor* for the high-frequency counterpart *hater*.

[3] Effect size (r) was calculated in line with Field (2018) where the F-ratio was converted to r. More specifically, the square root of [F] divided by [F + residual degrees of freedom (Error df)] was reported as large effect size.

4.3 Research question 3

The third research question aimed to explore the participants' performance in relation to task and individual word features before, during, and after completing the reading+replacing task. Although no specific hypotheses were made, task design and methodological factors were expected to influence the outcomes. This section summarizes participant performance for the word replacement task, reading comprehension, the survey of task interactiveness, the Vocabulary Size Test, and preknowledge and the posttest knowledge of the target words in varying depth. Participant profiles for the highest and lowest scores are also reported.

4.3.1 The word replacement task

Overall, for this task, Group A performed better and with more diversity (M = 8.38, SD = 1.835) than Group B (M = 7.96, SD = 1.400). More detailed distributions of scores can be seen in Figure 7.

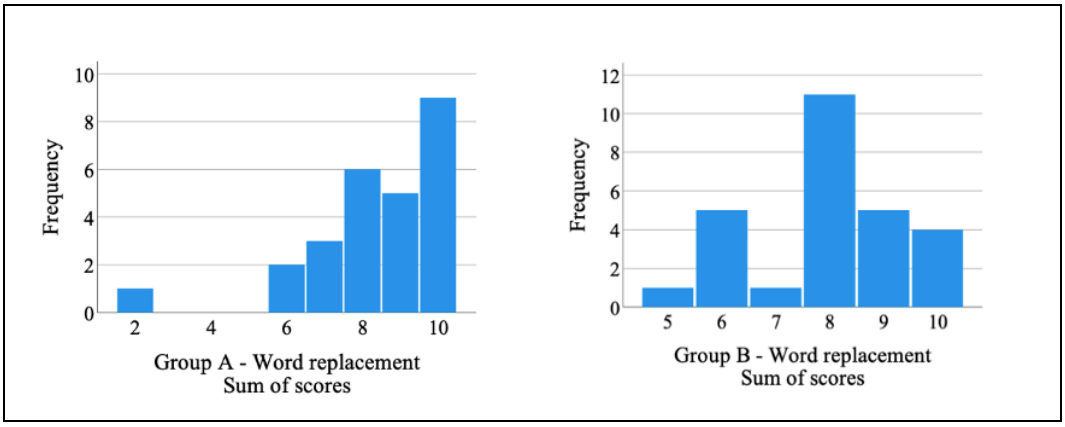

Figure 7. Group A's (left) and Group B's (right) word replacement task

To move from the word replacement task performance as a whole to the ratings of individual items, namely the target words, an overview of the number of correct and wrong replacements is presented in Table 9.

Table 9. Word Replacement Activity Number of Correct/Wrong Answers

Group	muddled		bungle		crass		dissection		confrontational		detractor		lament		premise		desecrate		ramification	
	C	W	C	W	C	W	C	W	C	W	C	W	C	W	C	W	C	W	C	W
A	24	2	14	12	23	2	22	4	24	1	25	1	24	2	21	4	17	9	24	2
B	27		12	15	21	6	25	2	22	5	26	1	27		18	9	11	16	26	1

Note. C = Correct, W = Wrong.

Three, but most prominently two, target words seemed to be replaced highly incorrectly in the groups: *bungle*, *premise*, and *desecrate*. Given the close similarity between the words *bungle* and *desecrate*, the participants seemed to miss the subtle difference between these words. The most frequent word replacements are listed below:

- The high-frequency counterpart *ruin* was wrongly replaced with *desecrate* by 11 participants (Group A), and similarly, its low-frequency counterpart *bungle* was replaced with *pollute* by 14 participants (Group B).
- Similarly, the high-frequency *pollute* was replaced with *bungle* by 5 participants (Group A), while the low-frequency *desecrate* was replaced with high-frequency *ruin* by 14 participants (Group B).
- The four participants (Group A) who incorrectly replaced the high-frequency *argument* with other words all chose different words (i.e., *ramification*, *confrontational*, *repository*, *instigate*), and for the low-frequency counterpart *premise*, there were some common wrong answers: *begin*, *container*, and *inactive* which were chosen by 3, 3, and 2 participants (Group B), respectively.

4.3.2 Reading comprehension

Five multiple-choice reading comprehension questions were asked to measure the participants' understanding of the text. Group A (M = 4.54, SD = .647) and Group B (M = 4.44, SD = .801) sum of scores were slightly different, with Group A getting a higher mean score overall. The results seemed to be favorable, hinting that overall the text was understood. Detailed summaries for both groups can be seen in Figure 8. Number of correct answers (i.e., 3, 4, and 5) and their frequencies are listed for each group.

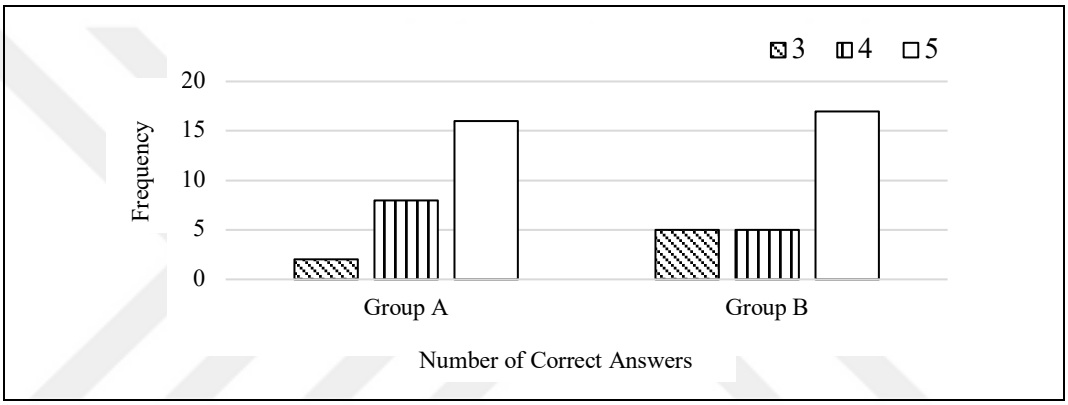

Figure 8. Reading comprehension performance by groups

4.3.3 Task interactiveness

Overall, Group A's ratings of task interactiveness (M = 17.85, SD = 3.107) seemed to be lower and more dispersed than those of Group B's (M = 18.59, SD = 2.485). Figure 9 displays group ratings for each item probing task interactiveness.

The first item measuring task interactiveness focused on participants' judgment of the suitability of the reading text for their proficiency level. Overall, highly favorable judgments were elicited from both Group A (M = 4.65, SD = .629) and Group B (M = 4.52, SD = .580), with relatively smaller dispersion compared to the rest of the items measuring task interactiveness.

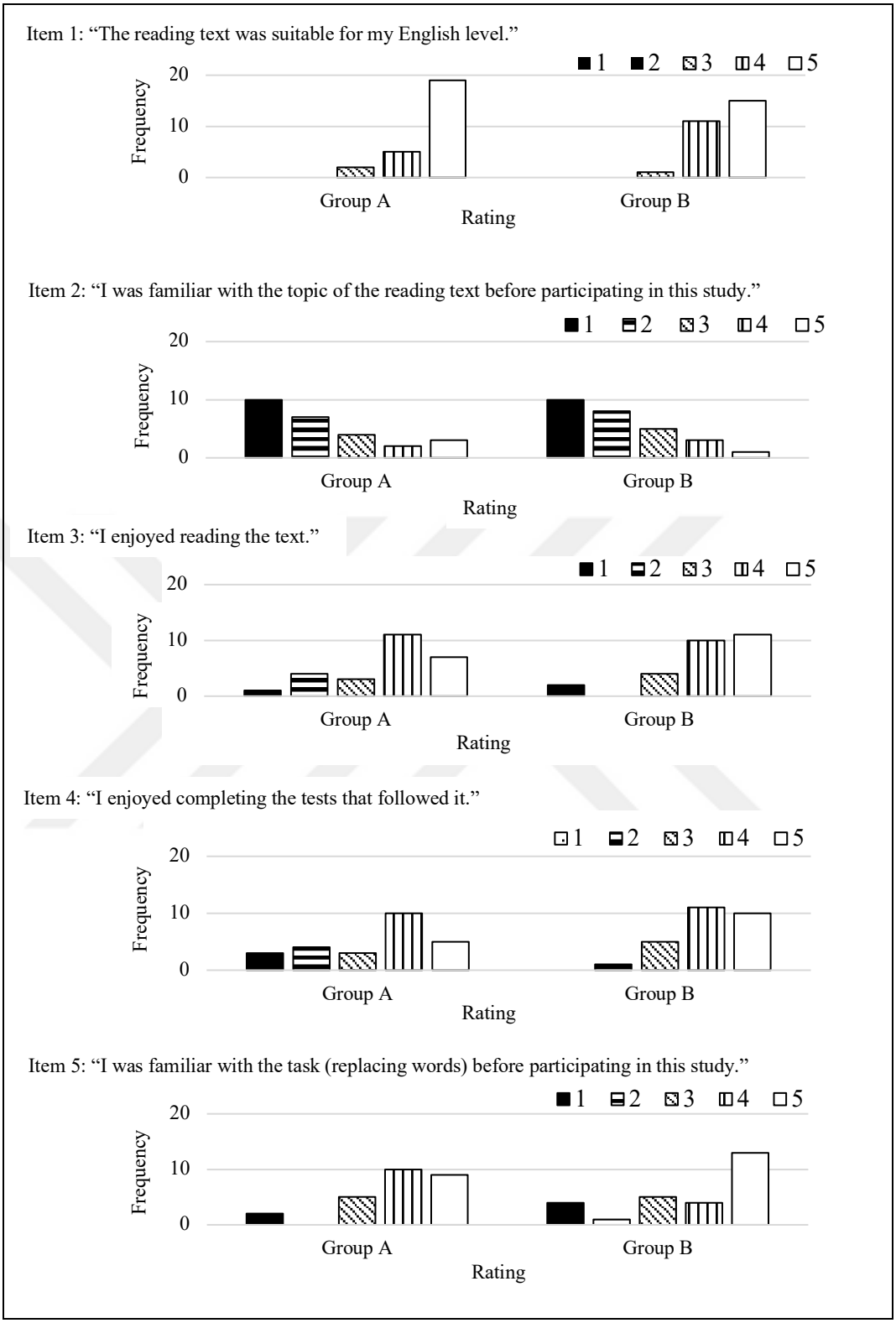

Figure 9. Participant ratings for task interactiveness

The second item focused on topic familiarity. This item was particularly noteworthy, with the overall lower mean rating scores when compared to the other items in its category, with Group A (M = 2.27, SD = 1.373) and Group B (M = 2.15, SD = 1.167) tending to disagree.

The third item concerned how appealing the reading texts were. The mean ratings seemed favorable, Group A (M = 3.73, SD = 1.151) and Group B (M = 4.04, SD = 1.126), but the ratings were highly dispersed.

The fourth item aimed to measure how appealing the tests were for the participants. For the tests that followed the reading text (i.e., the word replacement task and the reading comprehension task), the mean ratings of Group A (M = 3.40, SD = 1.323) and Group B (M = 4.11, SD = .847) seemed to be close to the favorable end of the scale, with the former group showing more diversity than the latter.

The final item probing task interactiveness examined task familiarity. The mean ratings of Group A (M = 3.92, SD = 1.129) and Group B (M = 3.78, SD = 1.476) showed a tendency to indicate familiarity with the word replacement task. Dispersion in the

ratings seemed to be quite visible, with Group B's standard deviation score being the highest among all of the task interactiveness items.

It would be important to draw attention to the items which elicited a rating of 1 and 2 combined, by at least 50% of the participants. For Group A, (65.4%) and for Group B (66.7%), item 2, (i.e., topic familiarity) was the only item with such combined ratings.

The measures of task interactiveness were the most favorable for *text suitability* and the lowest for *topic familiarity* both for groups. The task aspects between these were, in a decreasing order of mean scores, *task familiarity–text*

appeal–test appeal for Group A and *test appeal–text appeal–task familiarity* for Group B.

4.3.4 Vocabulary size scores

On average, Group A's vocabulary size (M = 8852.17, SD = 1557.349, *Min* = 6000, *Max* = 12200) was a little lower than that of Group B (M = 9033.33 , SD = 2199.539, *Min* = 5300, *Max* = 12700).

4.3.5 Group-based analyses

4.3.5.1 Preknowledge

Group A's (M = 3.73, SD = 2.089) and Group B's (M = 4.04, SD = 2.066) sum of preknowledge scores showed that all of the target words were not unknown by all of the participants. There seemed to be certain tendencies of the groups. To start with Group A, as displayed in Figure 10, the three words which were chosen least as unknown were, *confrontational* (N = 4), *premise* (N = 5), and *muddled* (N = 13). In contrast, *bungle* (N = 23), *crass* (N = 23), and *detractor* (N = 21) were the top three words chosen as unknown.

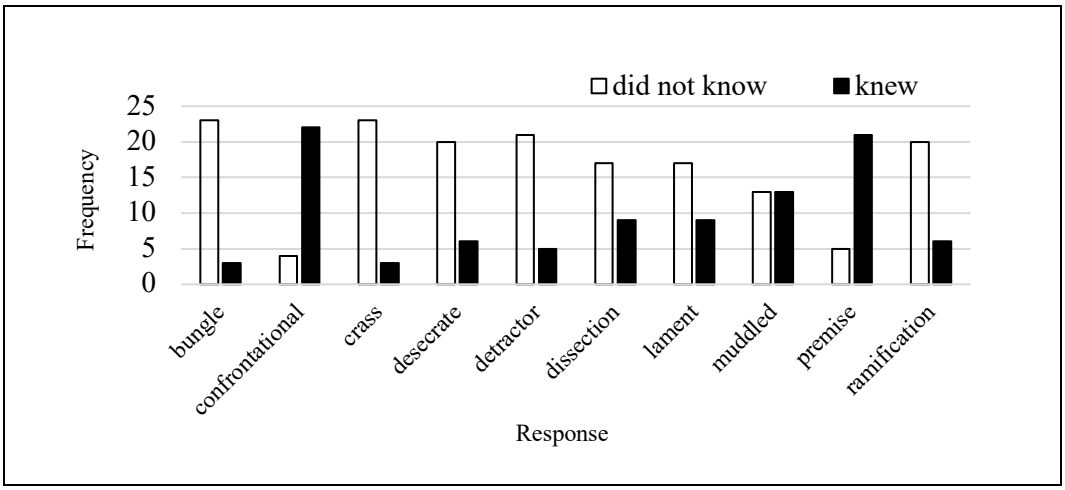

Figure 10. Group A's preknowledge of target words

As for Group B, the three words which were chosen least as unknown were *confrontational* (*N* = 3), *premise* (*N* = 7), and *muddled* (*N* = 11), as is shown in Figure 11. The top three words chosen as unknown were *crass* (*N* = 24), *detractor* (*N* = 24), and *bungle* (*N* = 22).

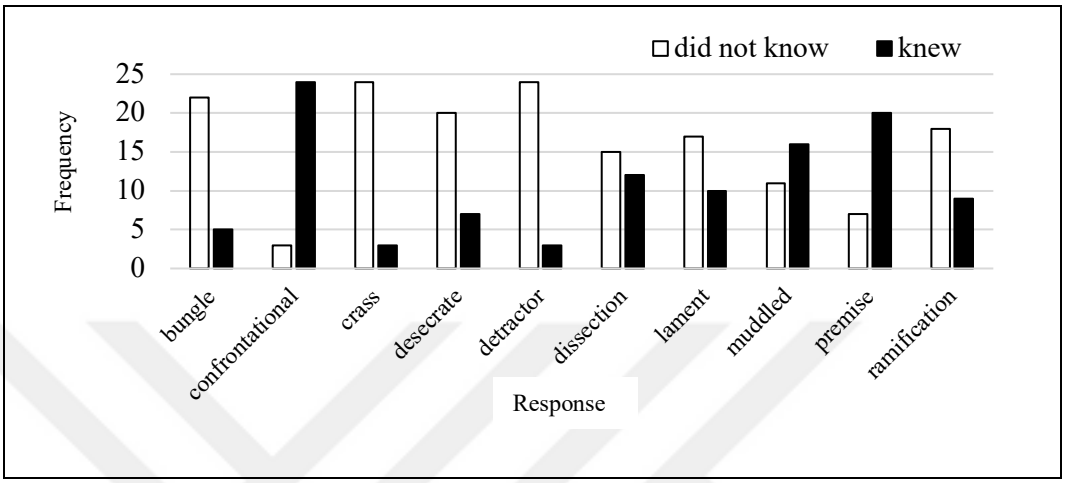

Figure 11. Group B's preknowledge of target words

As can be seen, in both groups, the same three words tended to get the fewest and most markings as unknown words, and the ideal scenario of zero knowledge could not be attained.

4.3.5.2 Posttest knowledge

Three target words with the lowest and the highest mean scores in each group were scrutinized. For Group A, as displayed in Figure 12, the lowest immediate posttest mean scores were for *lament* (*M* = 0.08), *crass* (*M* = 0.19), and *dissection* (*M* = 0.19) and the highest were for *muddled* (*M* = 0.68), *premise* (*M* = 0.64), and *confrontational* (*M* = 0.58). The lowest delayed posttest mean scores were for *bungle* (*M* = 0.09), *ramification* (*M* = 0.13), and *lament* (*M* = 0.13), whereas the highest average scores were for *premise* (*M* = 0.61), *muddled* (*M* = 0.57), and *confrontational* (*M* = 0.43).

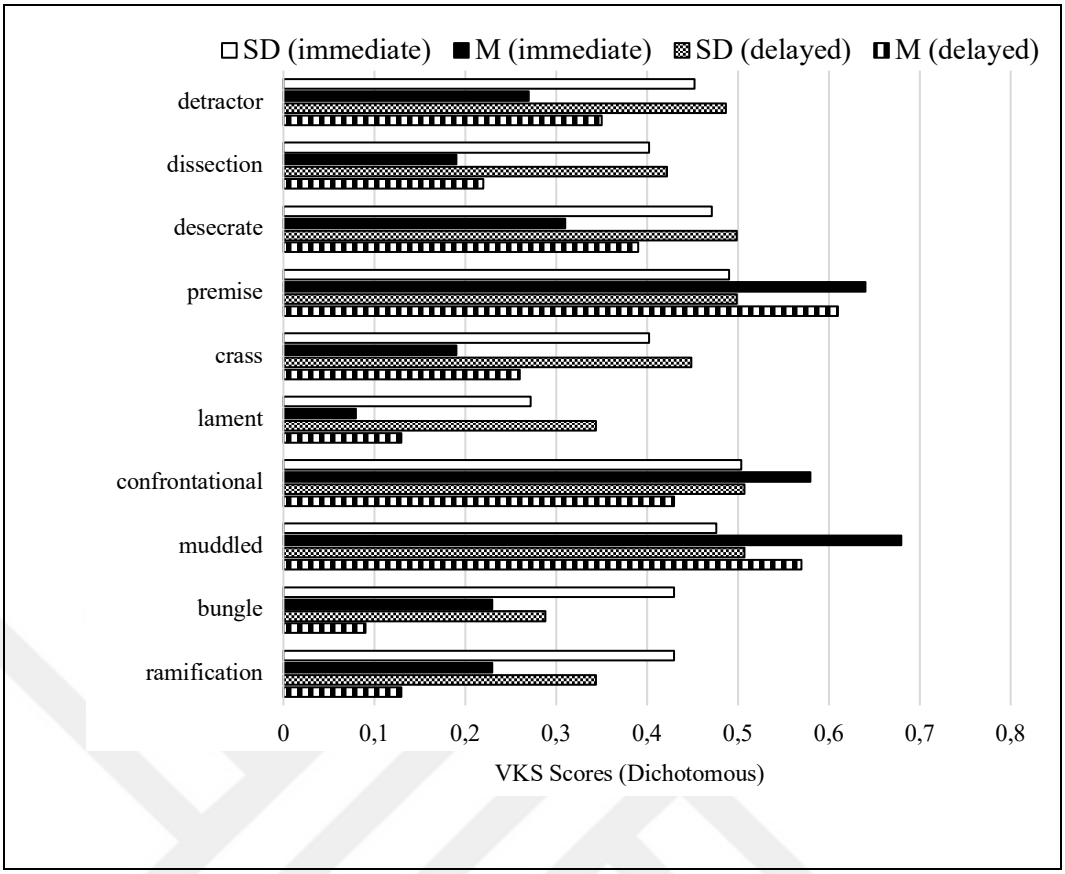

Figure 12. Posttest performance for Group A

As can be seen in Figure 13, for Group B, the lowest immediate posttest mean scores were for *bungle* (*M* = 0.22), *detractor* (*M* = 0.22), and *lament* (*M* = 0.22), whereas the highest were for *confrontational* (*M* = 0.76), *muddled* (*M* = 0.56), and *premise* (*M* = 0.44). The lowest delayed posttest mean scores were for *lament* (*M* = 0.08), *bungle* (*M* = 0.12), and *crass* (*M* = 0.2) and the highest were for *confrontational* (*M* = 0.52), *premise* (*M* = 0.52), and *dissection* (*M* = 0.4).

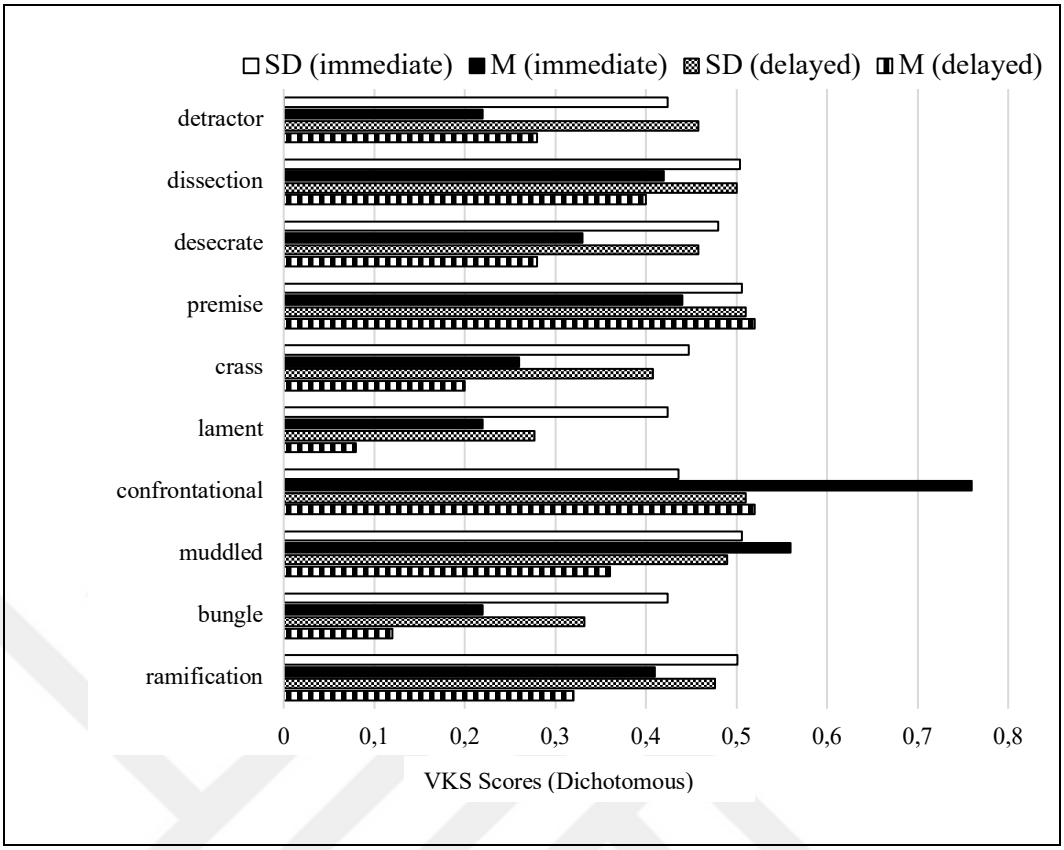

Figure 13. Posttest performance for Group B

There were some common words found for the three target words with the lowest and highest mean score categories in both posttests for both groups. For the immediate posttest lowest mean score category, it was *lament*, and for the highest mean category, the common words were *muddled*, *premise*, and *confrontational*. As for the delayed posttest, *bungle* and *lament* were the common target words with the lowest mean scores, and *premise* and *confrontational* were the words with the highest mean scores. To take these commonalities even further, the target words that retained their status in both posttests by both groups were *lament* (i.e., lowest mean score category) and *premise* and *confrontational* (i.e. highest mean score category). An interesting finding here concerned the word *lament*, a target word which did not make it to the top three previously unknown words in both groups. *Lament* appearing as the shared word between the groups as a word from the lowest mean score

category in both posttests might mean that its rival words were learned better or its marking as “unknown” was fewer than it should have been. The fact that some participants defined *lament* with one of its meanings which was not targeted in the present study and lost a score for this word was also a factor for the known-to-unknown status elicited.

4.3.5.3 Comparisons of pre- and posttest knowledge

In order to better portray the journey of the target words across the testing times, and also to come up with tangible patterns in retention or lack thereof, two more analyses were conducted. In these analyses, the three target words with the lowest/highest preknowledge ratings were at the core.

First, the three target words which were chosen least and most as unknown in the preknowledge test were compared with the target words receiving the lowest and highest mean scores in the posttests (i.e., the words shared by both groups, from among the three words ranking lowest and highest). From the words with the lowest preknowledge ratings, namely *bungle, crass*, and *detractor*, it was just *bungle* which appeared in the lowest category in the delayed posttest by both groups. From the words with the highest preknowledge ratings, namely *confrontational*, *premise*, and *muddled*, all the words appeared in the highest category for the immediate posttest and *premise* and *confrontational* stayed in the highest category in the delayed posttest.

Second, for the target words with the lowest preknowledge rates, the original scores on the VKS (i.e., scores of 1-5) were revisited. As discussed earlier, these words were *bungle*, *crass*, and *detractor*, each with over 20 markings as unknown in each group.

Figure 14 shows the percentages of VKS scores elicited for *bungle*. In Group A, the percentage of participants showed a decrease in the delayed posttest for score categories 1 (i.e., unfamiliar), 3 (i.e. meaning known), and 5 (i.e., semantically and grammatically correct use in a sentence) but not for 2 (i.e., familiar[4] but unknown). Indeed, no one got the full score 5 in the delayed posttest. As for Group B, similar trends were visible. The decline in scores for categories 1, 3, and 5 and the increase in the score category 2 was visible. Differently from Group A, score 5 category in the delayed posttest was not empty.

The score of 2 could be obtained in different ways, such as by simply choosing the category II in the scale (i.e., *"I have seen this word before, but I don't know what it means."*) or in the case of a wrong answer given to categories III, IV, or V. The meaning of score 2 is "*The word is familiar but its meaning is not known.*". Group A and B's performance suggested that this highly unknown word *bungle* mostly stayed familiar but unknown at the two different times it was measured. One decrease in percentages over time, namely in category I (i.e. "*The word is not familiar at all.*"), seemed to be a positive one because it signaled more familiarity with the target word.

[4] Familiarity was based on participants' recall of having seen a particular word before.

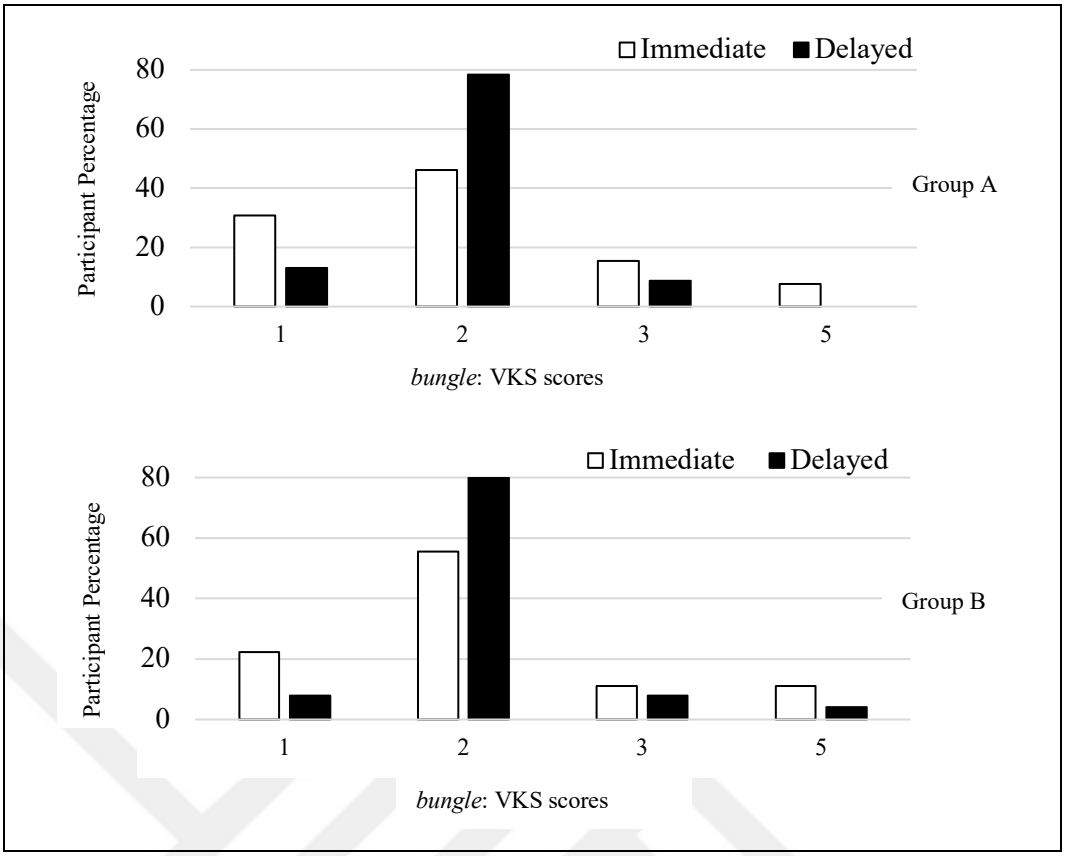

Figure 14. Group A's (top) and Group B's (bottom)VKS score percentages for

Figure 15 gives the percentage distributions of VKS scores for *crass*. For Group A, the decline over time in score categories 1 and 5 and the increase in categories 2 and 3 were visible. The increase in score category 3, which signifies mastery of word meaning, from the immediate to the delayed posttest seemed promising. This, however, was not the case for Group B. As is seen, there was an increase in category 2 over time, whereas the remaining categories revealed decline. In both groups, for the highly unknown word *crass*, a score of 2 was by far the most visible result for the delayed posttest. However, this was not the case for the immediate posttest where the percentage scores of 1 and 2 were either equal or close to each other. This might be an indicator of the fact that unlike *bungle*, *crass* initially tended to stay either (1) unfamiliar or (2) familiar but unknown, but in time it was mostly familiar but unknown.

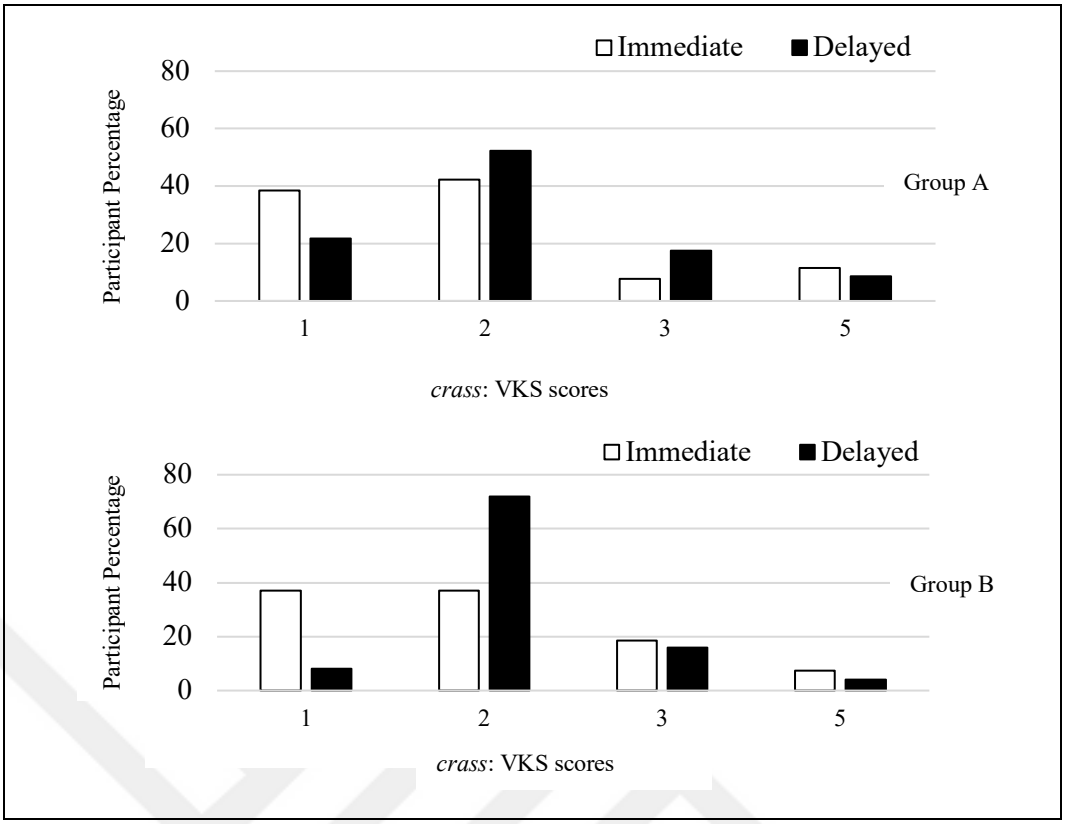

Figure 15. Group A's (top) and Group B's (bottom) VKS score percentages for

Finally, for *detractor*, the groups produced the score percentages shown in Figure 16. For Group A, as can be seen, score percentages for 1, 2, and 5 did not change drastically over time, though they slightly declined. The increase for the score of 3 was visible. The slight decline, or at least lack of an increase, in scores of 1 and 2 can be interpreted positively because the score (i.e. score of 5: "*The word is used with semantic appropriateness and grammatical accuracy in a sentence.*") showing the best knowledge of the target word attracted a substantial percentage of participants in the delayed posttest. A negative interpretation would focus on a lack of a clear transition from unfamiliar to familiar word meaning. Overall, at the two times it was tested, *detractor* tended to be either (1) unfamiliar or (2) familiar but unknown. The lack of sharp increase from score 1 to 2 found for *detractor* in Group A was a unique finding in this three-word set (including also *bungle* and *crass*). As for Group B's performance on *detractor*, the score category 1 showed the only and

the biggest decrease over time, whereas the remaining categories showed increase at different rates. The scores of 1 and 2 were the predominant scores during the time of the immediate posttest; however, over time the score of 2 became the most prominent one. In other words, while *detractor* was mostly either (1) unfamiliar or (2) familiar but unknown initially, it later was mostly familiar but unknown and there was also a trend for the mastery of meaning.

For *detractor*, both groups seemed to converge on the initial persistence of primarily (1) unfamiliar or (2) familiar but unknown scores elicited. Increasing trend in the score category 3, showing the mastery of word meaning, was visible.

To sum up, in all six analyses (3 words x 2 groups), the score category 2 was not surpassed by any other score categories at both times. There was just one case (*crass*, Group B) where the score category 1 was initially equal to category 2. Other times, the score category 2 was always the highest at both times. Except one case (*detractor*, Group A), there was an increase in score category 2 over time. This meant that these least known words, for the majority of the participants, stayed familiar but unknown. One improvement seemed to come from the score category 1, which showed a decrease over time in all six analyses. This showed that each target word was unfamiliar to smaller portions of participants over time.

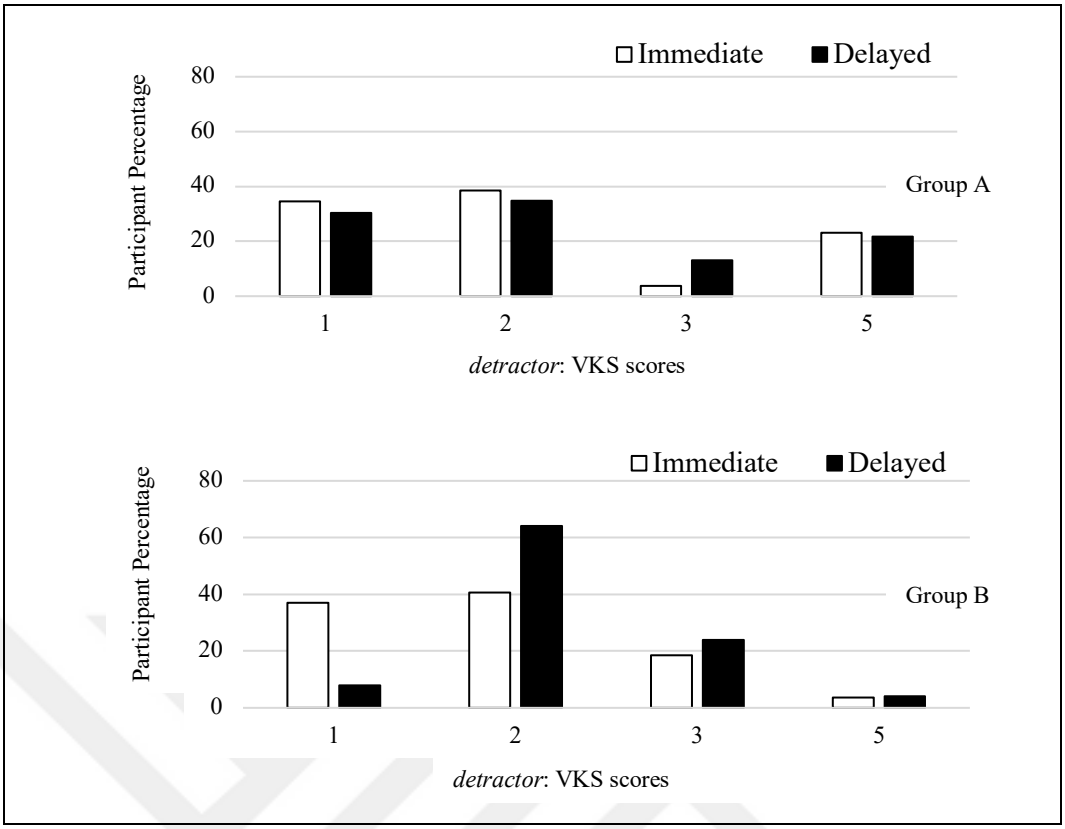

Figure 16. Group A's (top) and Group B's (bottom) VKS score percentages for *detractor*

4.3.5.4 Participant profiles for the highest and lowest scores

Each group's sum of scores for the (1) word replacement task, (2) survey of task interactiveness, (3) VST, (4) immediate VKS test, (5) delayed VKS test, (6) immediate VKS test (dichotomous), and (7) delayed VKS test (dichotomous) were further considered. Minimum three and maximum five participants were aimed for the highest and lowest scores; however, the different distribution of scores sometimes made it difficult to select the same numbers of participants for each of the seven tasks/tests. In Group A, there were two cases where the minimum-maximum number of participants range (i.e., 3 to 5 participants) was violated, with 6 participants (with the lowest delayed VKS test dichotomous[5] scores of 0 or 1) and 9 participants (with the highest word replacement task score of 10). In Group B, there

[5] Minimum and maximum scores were 0 and 10 respectively.

were four cases of 6 participants (with the lowest delayed VKS test dichotomous scores of 0 or 1, word replacement task score of 5 or 6, and task interactiveness score of 13, 14, or 16; with the highest immediate VKS test dichotomous scores of 6, 7, or 8). Participants whose scores were among the highest in at least four of the seven score categories are displayed in Table 10, with their performance details in the categories they ranked among the best.

Table 10. Participants with Highest Scores in at Least Four Score Categories

				VKS				
	VST	Word replacement	Task interactiveness	Immediate	Immediate (dichotomous)	Delayed	Delayed (dichotomous)	Preknowledge[a]
				Group A				
Participant #2	12000	10		39	7			8
Participant #15	10400		22			33	7	4
				Group B				
Participant #27	11800	10	22		6		7	8
Participant #35				42	8	41	7	3
Participant #47				32	6	28	8	7

[a]Preknowledge scores were included for comparison purposes only.

These participants, for whom the various tasks/tests yielded comparably more favorable results than the rest of the participants in their groups, did not excel in the same categories. In other words, it was not that the participants who had the largest vocabulary sizes, who replaced all the target words correctly, and who gave the highest ratings for task interactiveness also showed the highest retention.

Participants whose scores were among the lowest in at least four of the seven score categories are displayed in Table 11, with their performance details in the

categories they ranked among the lowest. For these participants with less favorable scores, there was no complete overlap across the categories. There was one commonality: None of the participants was among the lowest-scoring in the word replacement task. The relatively lower previous knowledge of the target words these participants had and the fine performance in the word replacement task were actually desirable parameters to measure the task effects in the present study. However, the resulting retention rates were quite unexpected.

In addition to the analyses of individual participants who tended to produce favorable and less desired scores across various tasks/tests, groups of participants with highest and lowest learning gains in each group were further analyzed and a profile for these participants was created. In other words, it was aimed to describe the vocabulary sizes, task performances, task perceptions, and previous target word knowledge of the participants with the highest and lowest retention scores. These descriptions naturally involved approximate numbers.

Table 11. Participants with Lowest Scores in at Least Four Score Categories

				VKS				
	VST	Word replacement	Task interactiveness	Immediate	Immediate (dichotomous)	Delayed	Delayed (dichotomous)	Preknowledge[a]
				Group A				
Participant #18				14	1	17	0	1
Participant #19	7500				1	19	1	2
				Group B				
Participant #53			16	17	1		1	3
Participant #49	7100		13			19	0	1
Participant #30	6600				1	20	0	3

[a]Preknowledge scores were included for comparison purposes only.

Firstly, for the immediate posttest, it was seen that the most successful participants were those who had the knowledge of 9,000 most frequent words families, replaced 8 words correctly, agreed that the tasks were interactive, and had preknowledge of 4 or 5 of the target words (see Table 12). The participants with the lowest rates of success were those who had the knowledge of 8,000-9,000 most frequent word families, replaced 7-8 words correctly, agreed the tasks were interactive, and had preknowledge of 1-2 of the target words.

Table 12. Participant Profiles for the Highest and Lowest Scores in the Immediate Posttest (Dichotomous)

		VST			Word replacement			Task interactiveness			Preknowledge		
	Number of participants	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max
					Group A								
top scores (6, 7)	4	9275	7200	12000	8.5	8	10	18	13	21	5	2	8
lowest scores (0, 1)	5	7933[a]	7500[a]	8500[a]	7.8	6	10	15.6	12	20	1.4	0	2
					Group B								
top scores (6, 7, 8)	6	9840[b]	7100[b]	11800[b]	8.8	8	10	18.5	13	22	4.1	1	8
lowest score (1)	4	9700	6600	12200	8.7	8	10	18.2	16	20	2.7	1	4

[a]There were two missing scores; therefore, reporting was based on three participants' scores. [b]There was a missing score; therefore, reporting was based on five participants' scores.

Secondly, for the delayed posttest, the most successful participants knew 9,000-10,000 most frequent word families, replaced 8 words correctly, agreed that the tasks were interactive, and knew 5 of the target words before the study (see Table 13). In contrast, the participants with the lowest rates of success had the knowledge of 7,000-8,000 most frequent word families, replaced 7-8 words correctly, agreed that the tasks were interactive, and had preknowledge of 1-2 of the target words.

Table 13. Participant Profiles for the Highest and Lowest Scores in the Delayed Posttest (Dichotomous)

		VST			Word replacement			Task interactiveness			Preknowledge		
	Number of participants	M	Min	Max	M	Min	Max	M	Min	Max	M	Min	Max
					Group A								
top scores (5, 6, 7)	5	9380	8300	10400	8.4	6	10	18.6	13	22	5	4	6
lowest scores (0, 1)	6	7616	6000	8500	8.5	7	10	16.5	12	20	1.6	0	4
					Group B								
top scores (6, 7, 8)	4	10600	9200	11800	8.2	6	10	18	14	22	5.7	3	8
lowest scores (0, 1)	6	8216	6600	12200	7.8	6	10	17.1	13	20	2.1	1	3

4.3.6 Analyses on the entire sample

The different groups in the present study did not behave significantly differently from each other. For this reason, group-based analyses were kept as brief as possible. More detailed analyses including the entire sample were carried out in order to see the status of the target words across the three times they were tested.

Checking the immediate and delayed posttest score pairs (i.e., the original VKS scores 1-5) for each participant was important to see the overall picture in target word retention. This check was twofold. First, to examine the possible fluctuations in the participants' knowledge of the target words as measured in the immediate and delayed posttests, the scores gained in the immediate posttest for each target word were subtracted from the scores gained in the delayed posttest (i.e., delayed posttest score – immediate posttest score). The resulting scores indicated both the extent of change and also whether that change was a progress or decline. For example, an immediate posttest score of 5 and a delayed posttest score of 2 would

give a score of -3 (i.e., delayed – immediate = 2 – 5 = -3), indicating a decline in word knowledge from immediate to delayed posttest, hence the minus sign. A score of 2 in both tests would give a score of 0, meaning that knowledge stayed stable. Finally, a score of 1 in the immediate posttest and 2 in the delayed posttest would give a score of 1 (i.e., delayed – immediate = 2 – 1 = 1), which would be an indicator of progress. Samples for these score analyses came from groups of 46, 47, or 48. Second, all score pairs (immediate posttest score*delayed posttest score) were documented for each target word. These two analyses concerned (1) progress and decline in VKS scores over time, and (2) the actual scores on which such progress and decline were based on, respectively. With the former analysis, the aim was to describe the bigger picture for each target word, and with the latter details were provided.

To start with *ramification*, 29 participants retained their exact score across the posttests (see Figure 17). However, the majority of these score pairs were scores of 2*2 ($N = 20$), which meant that *ramification* remained familiar but unknown. Losing or gaining 1 point was the next striking finding for this word, as was the case with 7 and 6 participants, respectively: Those whose knowledge deteriorated either lost their familiarity with *ramification* (2*1, $N = 2$) or maintained familiarity but forgot its meaning (3*2, $N = 5$), and those who showed progress gained familiarity (1*2, $N = 6$).

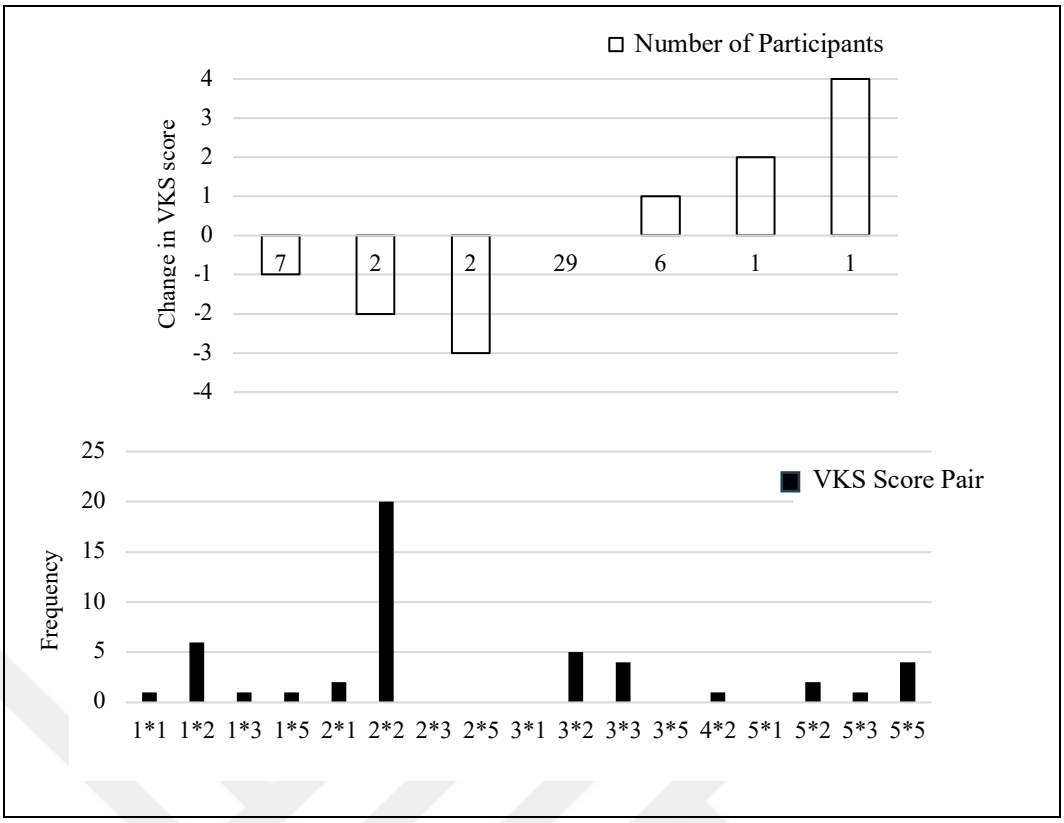

Figure 17. *Ramification:* Changes in posttest scores (top) and posttest scores (bottom)

As for *bungle*, 27 participants gained the same score in the posttests (see Figure 18). These consistent scores were mostly scores of 2*2 ($N = 23$), signaling that *bungle* was familiar but unknown. The tendency following this involved the cases of gaining and losing 1 point by 11 and 6 participants, respectively: Those who showed progress either gained familiarity (1*2, $N = 9$) or came to learn word meaning (2*3, $N = 2$), and those whose knowledge deteriorated either maintained familiarity but forgot word meaning (3*2, $N = 4$) or lost their familiarity with *bungle* (2*1, $N = 2$).

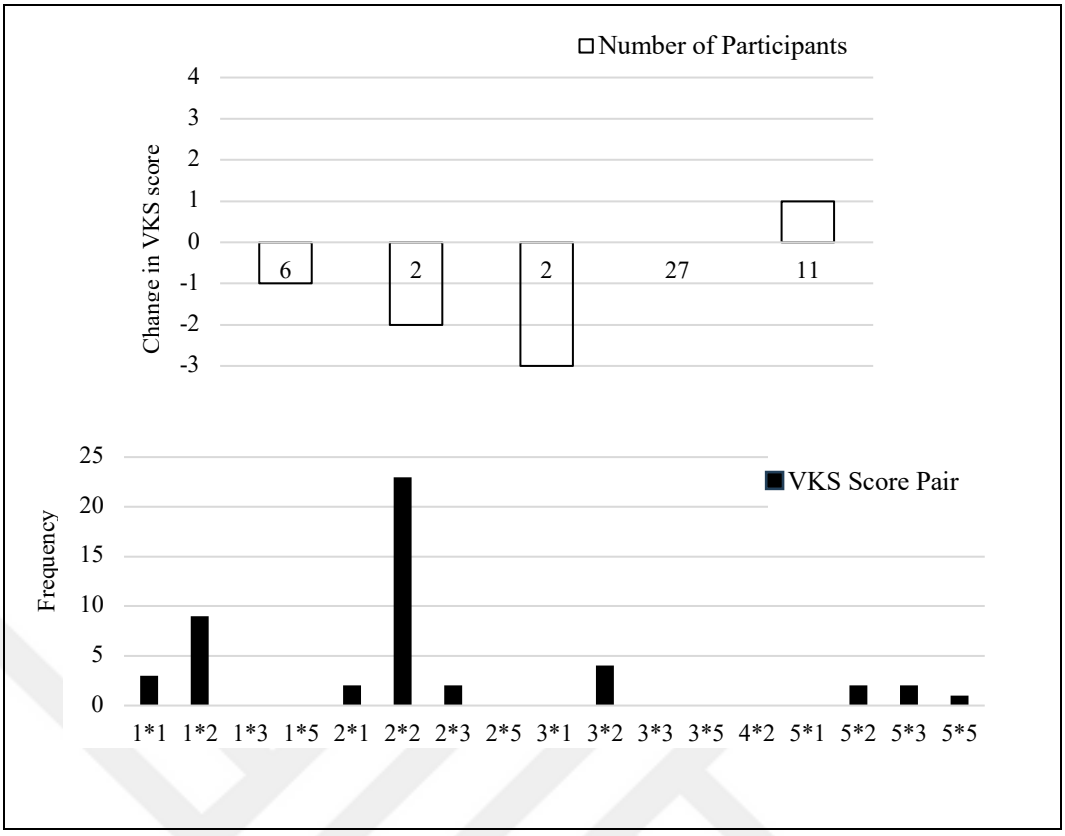

Figure 18. *Bungle:* Changes in posttest scores (top) and posttest scores

The degree of knowledge for *muddled* stayed the same for 25 participants (see Figure 19). Although the score pair of 2*2 was the most dominant ($N = 10$), pairs of 3*3 ($N = 8$), and 5*5 ($N = 6$) were quite competitive. This showed that *muddled* remained familiar but unknown ($N = 10$), but also known to a varying extent ($N = 14$). Losing 1 point was another visible pattern, with 9 participants showing decline in their knowledge at the delayed posttest. This was in the manner of score pair 3*2, with the participants maintaining familiarity but forgetting word meaning.

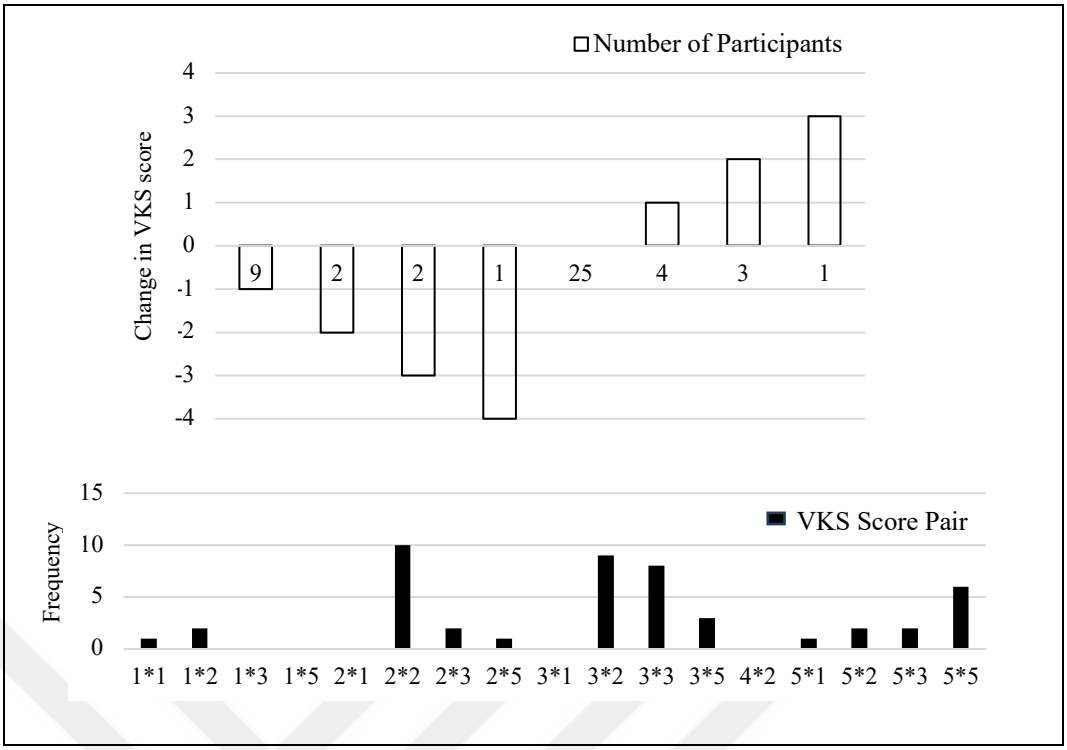

Figure 19. *Muddled:* Changes in posttest scores (top) and posttest scores

Confrontational seemed to yield one of the most dynamic sets of scores among the target words (see Figure 20). While 19 participants kept their scores across the posttests, tangible chunks of participants showed decline. In other words, 7, 7, and 6 participants lost 1, 2, and 3 points respectively. More specifically, the dominant identical score pairs of 2*2 ($N = 8$) and 5*5 ($N = 8$) showed that *confrontational* remained familiar but unknown and known quite well, respectively. Those who lost scores primarily retained word meaning but failed to demonstrate higher-level word knowledge (5*3, $N = 7$), lost higher-level word knowledge but maintained familiarity (5*2, $N = 6$), and maintained familiarity but forgot word meaning (3*2, $N = 6$).

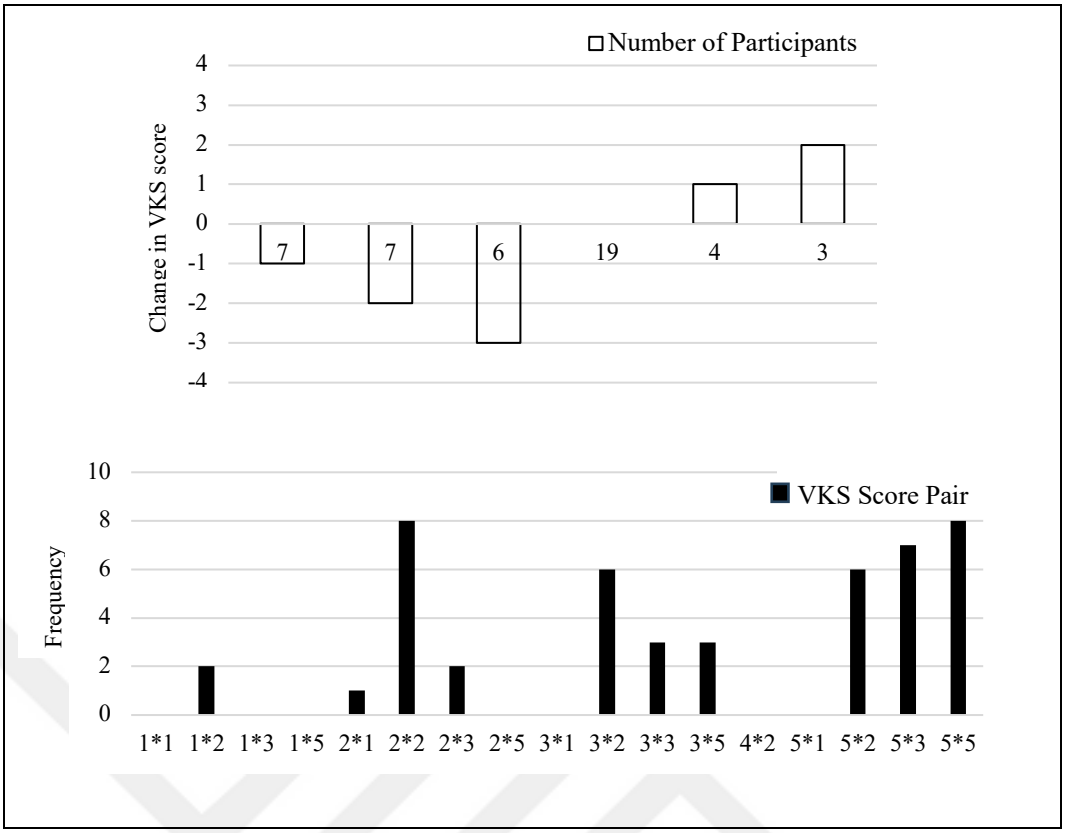

Figure 20. *Confrontational:* Changes in posttest scores (top) and posttest scores (bottom)

Lament seemed to produce rather conservative results, with 32 participants retaining their knowledge (see Figure 21). The predominant score pair was 2*2 (*N* = 30), showing that *lament* remained familiar but unknown. The next salient finding was the 1-point progress achieved by 8 participants, where the dominant score pair 1*2 (*N* = 7) indicated gained familiarity with the target word.

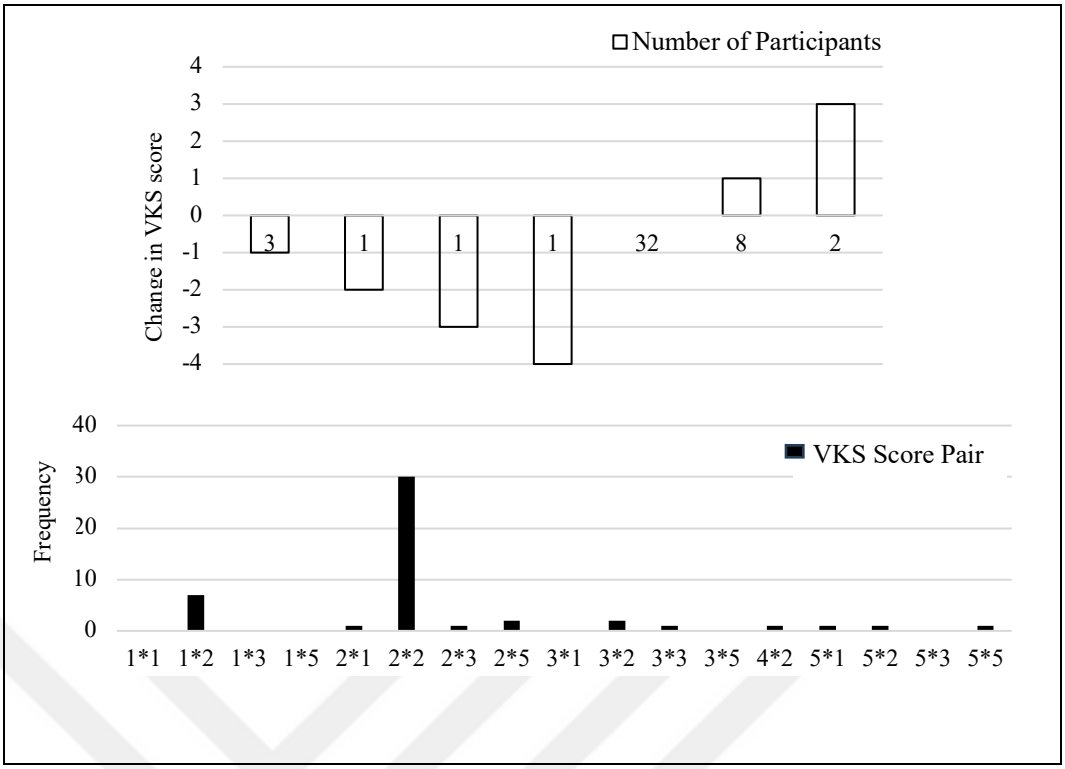

Figure 21. *Lament:* Changes in posttest scores (top) and posttest scores

The changes observed for *crass* did not involve extremes; the participants either lost or gained maximum 2 points (see Figure 22). Following the 25 participants who kept their scores the same were 11 participants who increased their knowledge by 1 point as well as 7 participants whose knowledge showed a 1-point decline. Those who retained their exact scores mostly belonged to the 2*2 group (N = 16) for whom *crass* remained familiar but unknown. For progress, the dominant score pair was 1*2 (N = 9), showing gained familiarity. For decline, the score pair 3*2 (N = 5) was more prominent, with the participants maintaining familiarity but forgetting word meaning.

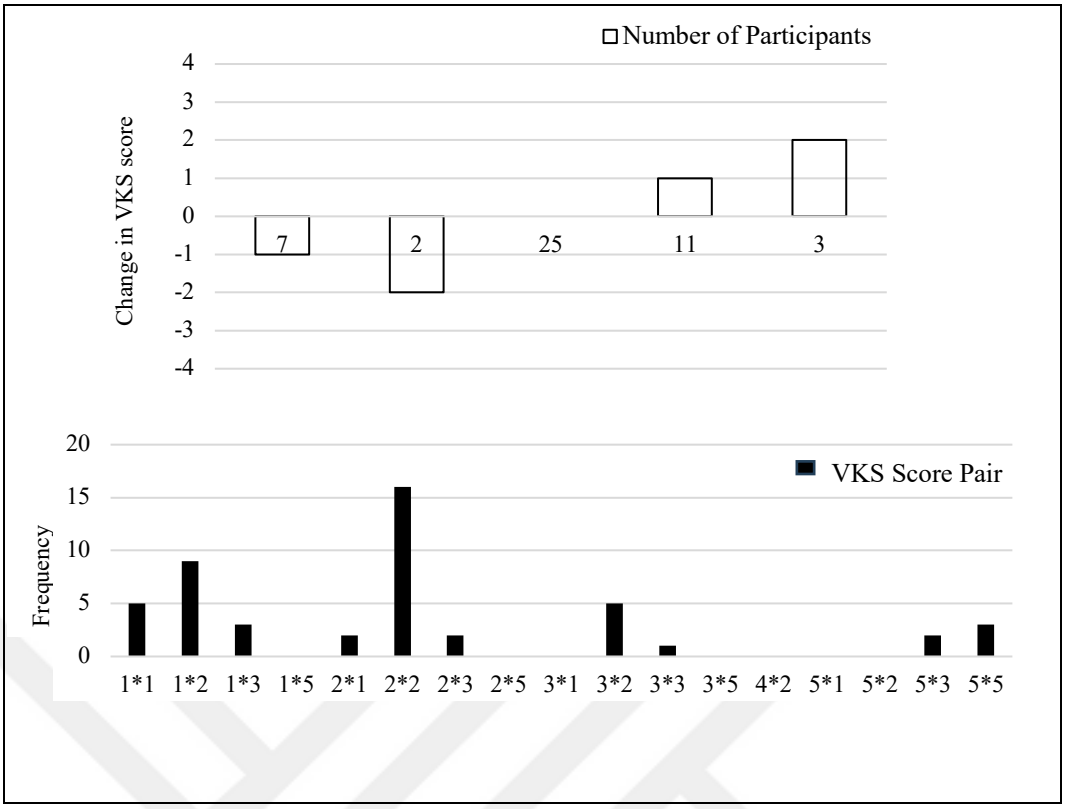

Figure 22. *Crass:* Changes in posttest scores (top) and posttest scores (bottom)

Premise was another word for which a large portion of scores, i.e., 31, showed no change across the posttests (see Figure 23). The dominant identical score pair was 2*2 ($N = 16$), showing that *premise* remained familiar but unknown. The following pairs 3*3 ($N = 7$) and 5*5 ($N = 8$) showed that the target word remained known to a varying extent. The next noteworthy performances were by 6 participants who ended up with a 2-point decrease in knowledge and another 6 who showed a 1-point improvement. Those whose knowledge deteriorated retained word meaning but failed to demonstrate higher-level word knowledge (5*3, $N = 6$). Improvement was visible for the 2*3 pair ($N = 4$), with the mastery of word meaning.

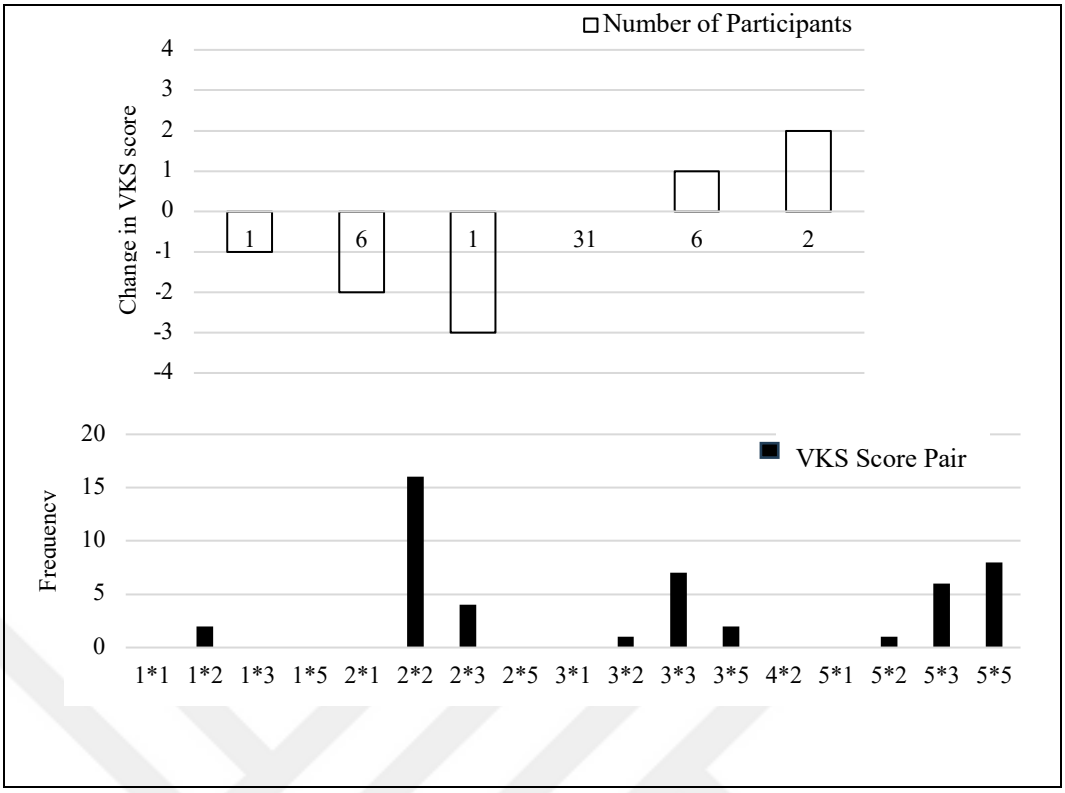

Figure 23. *Premise:* Changes in posttest scores (top) and posttest scores

Desecrate stood out with the highly noticeable 1-point increase in participants' knowledge in the delayed posttest (see Figure 24). More specifically, 18 participants managed to show progress from the immediate to the delayed posttest. The majority of progress came from the score pair 1*2 ($N = 13$) and meant gained familiarity with *desecrate*. The number of participants whose knowledge remained the same was 22, with the dominant score pair 2*2 ($N = 11$) signaling that *desecrate* remained familiar but unknown.

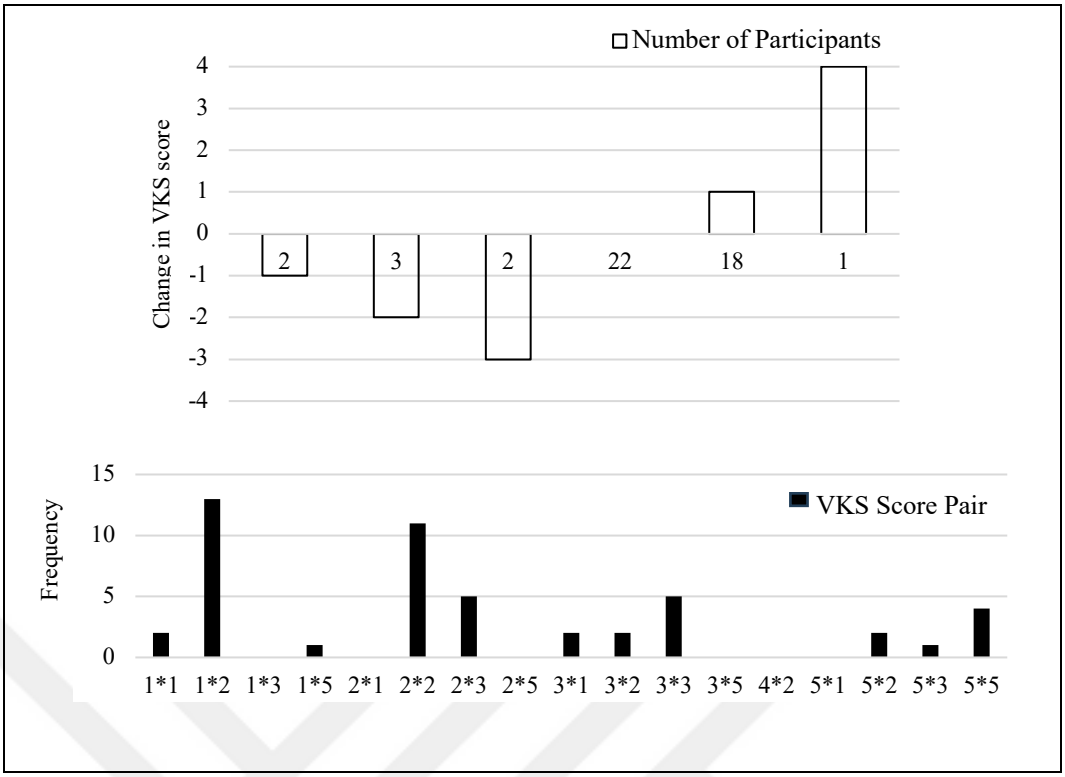

Figure 24. *Desecrate*: Changes in posttest scores (top) and posttest scores

Dissection yielded a trend similar to that of desecrate (see Figure 25). The largest portion of scores showing no change in word knowledge was followed by a 1-point increase in knowledge. More specifically, while 23 participants kept their scores, 11 participants increased their scores by 1 point. The majority of the identical scores were in the form of 2*2 ($N = 16$), implying that *dissection* remained familiar but unknown. Progress came from the score pairs 1*2 ($N = 6$) and 2*3 ($N = 5$), which translated to gained familiarity with the target word and the mastery of word meaning, respectively.

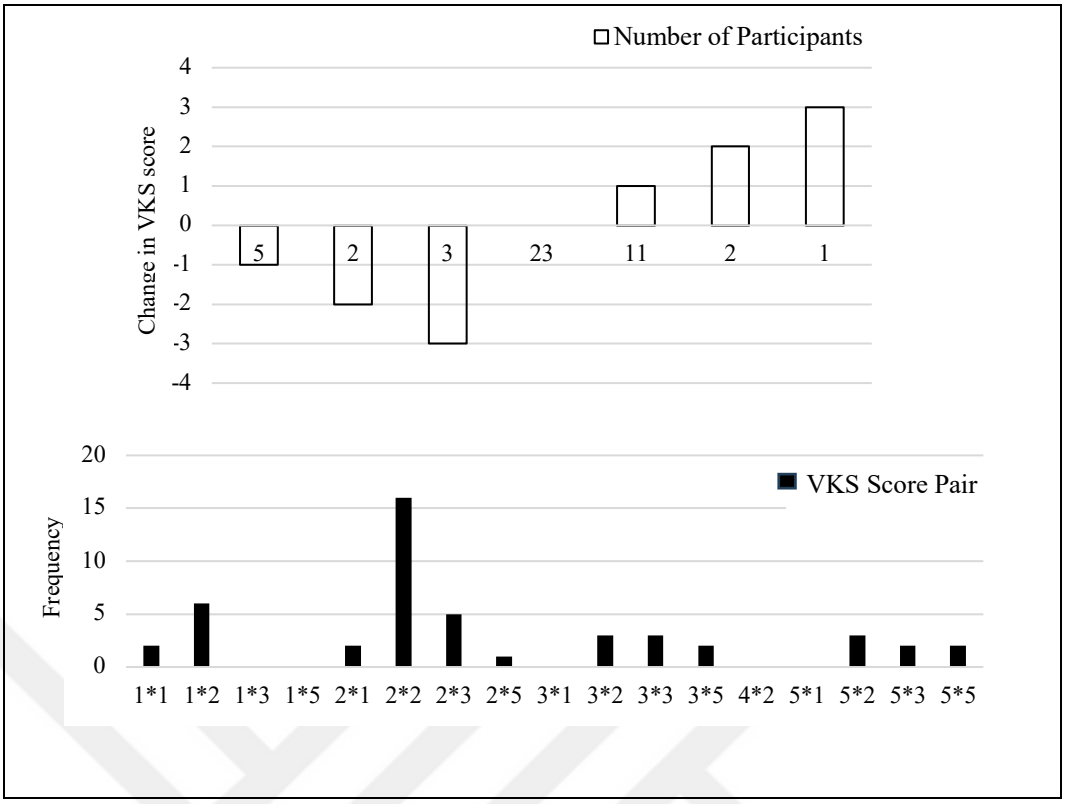

Figure 25. *Dissection:* Changes in posttest scores (top) and posttest scores

Similar to dissection and desecrate, *detractor* showed the most salient trends in relation to scores with no change and scores that increased by 1 point (see Figure 26). While 22 participants kept their scores, 13 participants showed a progress of 1 point. Those keeping their scores mainly belonged to the score pair 2*2 ($N = 13$) and *detractor* remained familiar but unknown to them. Progress was dominated by the score pair 1*2 ($N = 10$), showing gained familiarity with the target word.

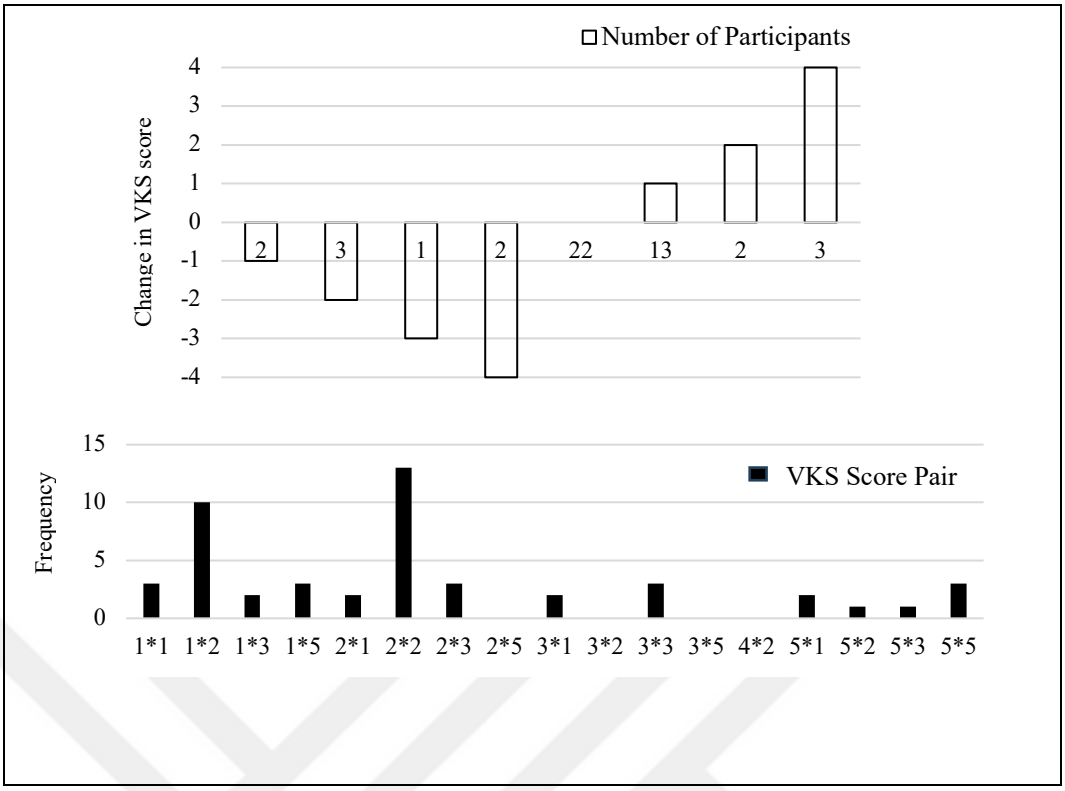

Figure 26. *Detractor:* Changes in posttest scores (top) and posttest scores

After the analyses of individual target words, an overall examination of the score pairs elicited by the participants was conducted (see Figure 27). It was seen that the predominant tendency for a target word was to get a score of 2 at the time of both posttests (i.e., 2*2). This meant that the target word remained familiar but unknown. This tendency was followed by getting a score of 1 in the immediate and 2 in the delayed posttest (i.e., 1*2). This signified a transition from a word being unfamiliar to familiar, with meaning still being unknown. The third dominant performance was getting top scores in both posttests (i.e., 5*5), which meant that the word remained known and its semantically and grammatically correct use in a sentence was preserved. For the scores which remained the same across the posttests, the target words remained familiar but unknown (i.e., 2*2, $N = 163$), known quite well (i.e., 5*5, $N = 40$), known (i.e., 3*3, $N = 35$), and unfamiliar (i.e., 1*1, $N = 17$), in a decreasing order.

The top three patterns of progress came for gained familiarity (i.e., 1*2, $N = 66$), mastery of meaning (i.e., 2*3, $N = 26$), and higher-level word knowledge attainment (i.e., 3*5, $N = 10$). The top three patterns of decline showed that the participants maintained familiarity with the target word but forgot its meaning (i.e., 3*2, $N = 37$), retained meaning but failed to demonstrate higher-level word knowledge (i.e., 5*3, $N = 24$), and maintained familiarity but lost higher-level word knowledge (i.e., 5*2, $N = 20$).

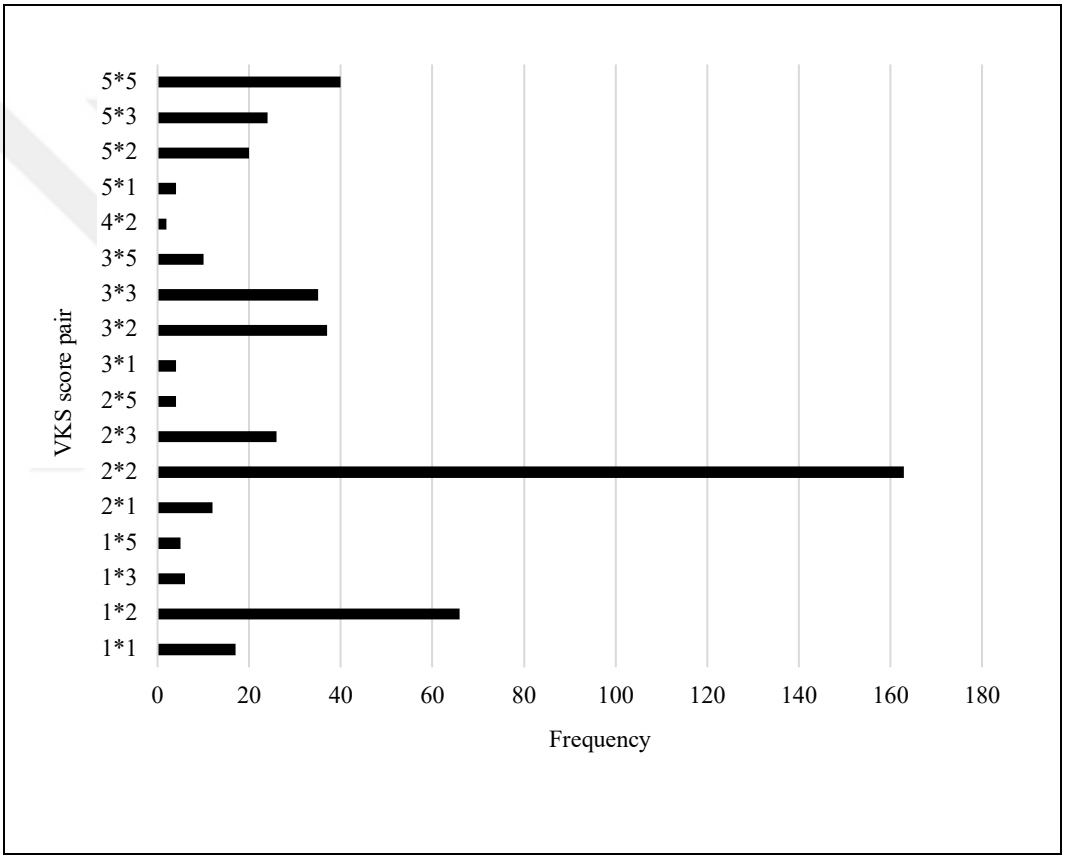

Figure 27. Total number of cases for immediate*delayed posttest scores

For each target word, the number of participants who showed progress and decline in their scores across the posttests was also calculated (see Figure 28). Overall, it seemed that *detractor*, *dissection*, *desecrate*, *crass*, *lament*, and *bungle* were the target words for which more participants showed progress than decline in their word knowledge. *Premise* was associated with an equal number of participants

with progress and decline scores. *Confrontational*, *muddled*, and *ramification*, on the other hand, elicited more cases of decline than progress.

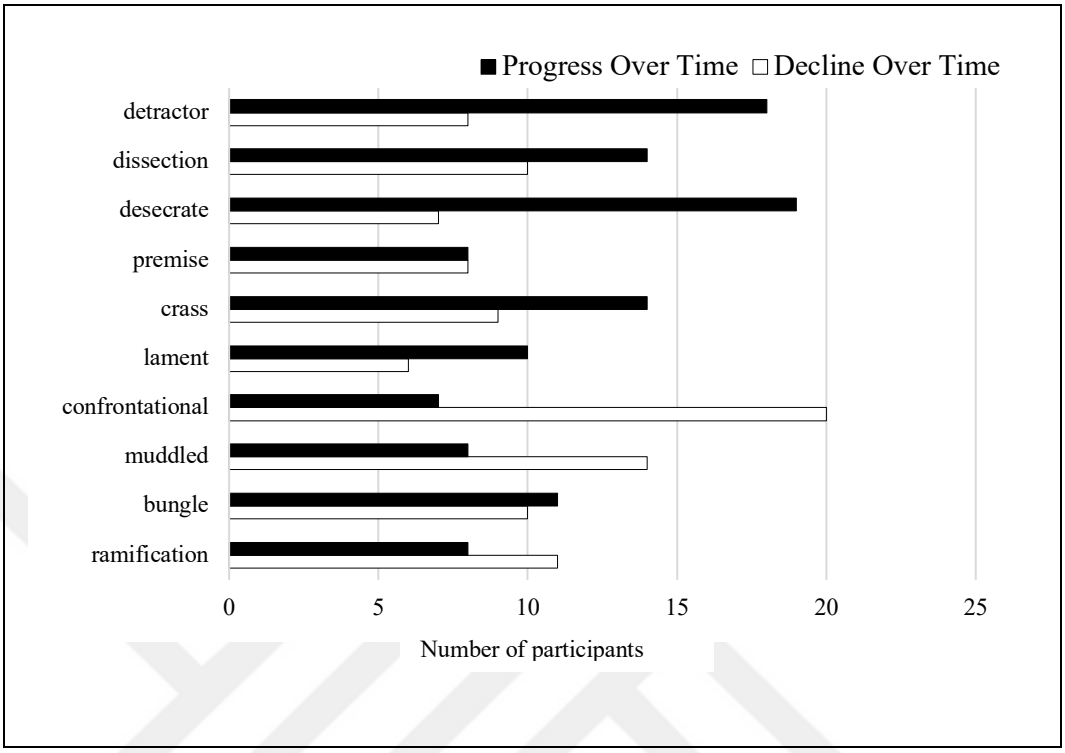

Figure 28. Number of participants with progress and decline in scores across posttests

Furthermore, by adding together all the decline and progress scores for each target word, the magnitudes of such changes were calculated (see Figure 29). Overall, *detractor*, *desecrate*, *crass*, and *lament* were the target words for which scores gained were more than scores lost. *Dissection* lost the same amount of score as it earned. *Ramification*, *bungle*, *muddled*, *confrontational*, and *premise* lost more scores than they gained.

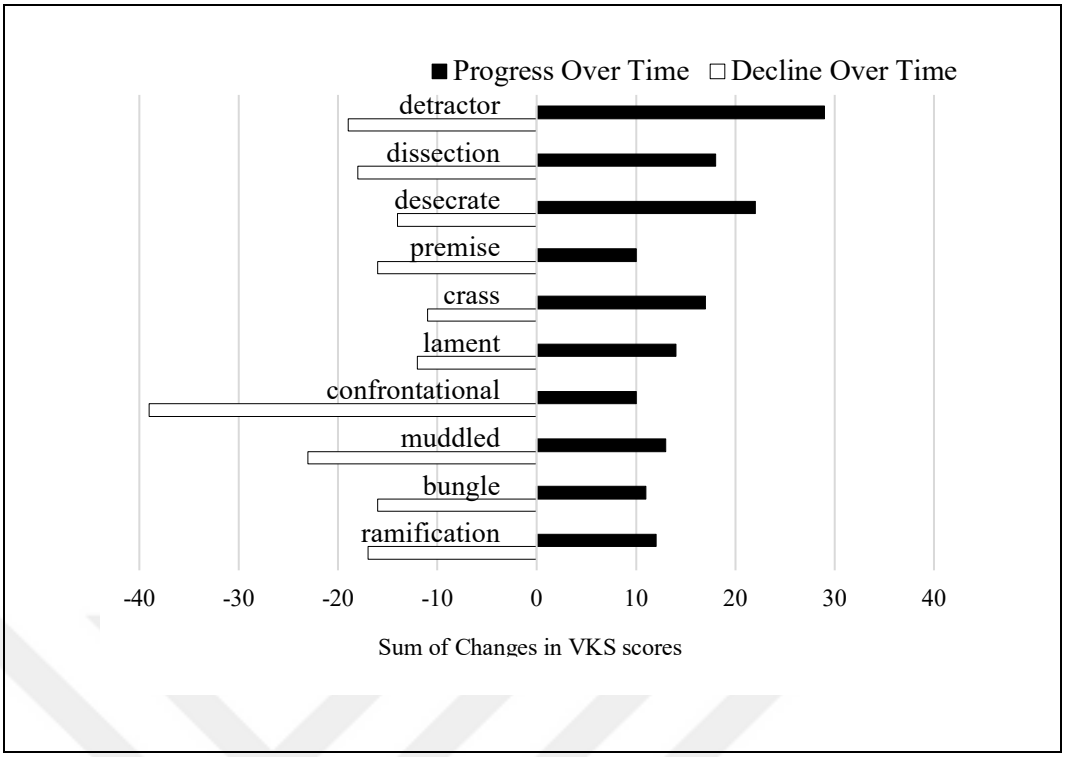

Figure 29. Magnitude of progress and decline per each word across posttests

In summary, the results showed no significant main effect of group. This meant that replacing the low-frequency words in the text with their high-frequency counterparts and vice versa did not result in significantly different gains in target word knowledge. There was a significant main effect of time, with the participants' preknowledge of the target words being significantly higher than their knowledge measured at the end of the study (i.e., in the delayed posttest). In other words, it seemed that around two weeks after engaging in the word replacement task, the participants' knowledge of the target words deteriorated, let alone staying stable. There were no significant effects of time x group interaction. As for the recall of target word forms, the response ratings were far from ideal, with highly low numbers of participants coming up with the target words in appropriate contexts.

To conclude, certain task components did not work in their intended ways. Therefore, it became really difficult to argue, with due precision and confidence, that the answers found for the research questions were not flawed. This chapter was

limited to the answers to the research questions, and further issues will be discussed in the following chapter.

CHAPTER 5

DISCUSSION

The present study compared two tasks which were identical in terms of *need* (i.e., moderate) and *search* (i.e., absent) dimensions but different in terms of *evaluation* (i.e., moderate vs. absent). As such, a task with a higher involvement load index, an index of 2, was compared with a task with an index of 1. The task with the higher involvement index, which required the use of the low-frequency words in the word list and their replacement with high-frequency counterparts in the text, was expected to result in more learning gains than the task with the lower involvement index, which required the same replacement in the reverse direction. This was not the case, and the tasks did not elicit different learning gains, which failed to support Hypothesis 1. The participants' knowledge of the target words did not increase, statistically or otherwise, as a result of completing the tasks. In fact, steady decreases were observed across testing times, among which the long-term retention was significantly lower than preknowledge.

For the relative position of each target word in the participants' performance, a few caveats were worth making. The most frequent word family represented among the low-frequency target words was the 3k family, with two words *confrontational* and *premise*. In an analysis of the words with the highest mean scores for both groups and for both posttests, these words were encountered. This meant that, on average, *confrontational* and *premise* elicited the best performance in the present study. It is difficult to argue that the participants learned these words as a result of engaging in the tasks, because these words were already in a good place at the start of the study. The word family information here proved its usefulness. It

should, however, be kept in mind that this discussion is based on mean scores and words in relation to each other. Although these words were in a comparably better position than the others, they cannot be associated with complete mastery by each and every participant.

The following word family was 5k, with the word *lament*. It turned out to be the only word found in an analysis of the words with the lowest mean scores for both groups and for both posttests. On average, *lament* was the word with the lowest success rate in this study. This might have been partly due to the other meanings it has, which were considered incorrect when provided by the participants.

To turn from a top-and-bottom analysis to a word-based inspection, the popularly previously unknown trio of *crass*, *detractor*, and *bungle* can be argued to remain modest. In other words, none of them ever made it to the top four mean scores in any of the four cases (i.e., two groups x two posttests). The finding that these words tended to remain unknown was not surprising; however, the cases where the participants claimed unfamiliarity were quite unexpected. Marking a target word unfamiliar in the immediate posttest could be explained in many ways: The participant at least encountered (not necessarily read its definition or replaced it correctly) the word during the task completion stage but forgot it later in the same session, or the target word was never encountered during the task completion stage. *Desecrate*, in three of the cases, was learned better than this trio. This paralleled the runner-up status it had after these target words with the lowest preknowledge ratings. *Muddled* was in top two in three of the cases, and elicited better performance than its fellow 7k word family members *dissection* and *detractor* in three of the cases. *Muddled*, therefore, seemed to maintain the superiority it had over *dissection* and *detractor* in preknowledge ratings. This might, again, have been a case of target

words keeping their relative positions among themselves before and after participating in the study. *Ramification*, did not make it to the top four in all cases. Finally, *dissection* always elicited better performance than *lament* did.

Although it might be rather difficult to make true comparisons between this study and the others reviewed in the previous sections, mainly due to study components (e.g., task types, tests used, etc.), some grounds for comparison seem to be existent.

The first point of comparison is naturally the overall involvement indexes of tasks and how they are reflected in the results: as support for the ILH or not? It should be kept in mind, though, such a comparison across different studies will probably be very simplistic. Comparisons across studies with very similar instruments would be more insightful. The present study did not lend support to the ILH. In the literature reviewed, the studies (i.e., Eckerth & Tavakoli, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008a) including a comparison between two tasks with the same *need*, *search* and *evaluation* indexes as the present study (i.e., +, -, + = 2 vs. +, -, - =1) seemed to be good candidates for further comparison. There were cases of significantly better performance, initially or later, elicited from the task with the higher involvement index in all of these four studies. Yet, there were also findings which pointed to a lack of significant difference between the tasks with the aforementioned involvement indexes (e.g., Eckerth and Tavakoli, 2012; Hulstijn & Laufer, 2001; Kim, 2008a). As can be seen, even in a single study, there are findings that support the ILH and those that do not. Of course, such findings can be discussed on many grounds such as the times when the target words were tested, target word properties, and task properties. The reading texts used in these studies were in the form of expository texts which were adapted from an English for Academic Purposes

coursebook (Eckerth & Tavakoli, 2012), a description adapted from an encyclopedia (Keating, 2008), a letter to editor from a national reading comprehension exam (Hulstijn & Laufer, 2001), and a text adapted from a coursebook (Kim, 2008a). Indeed, it seems to be that the very tasks in the present study make it difficult to maintain comparisons with the tasks which likewise failed to support the ILH. Although the LowText and the HighText tasks can be seen essentially as gap-filling tasks, which seem to be quite common among the studies reviewed, they are not. They are not categorically different tasks themselves either. Rather, they are just two slightly different versions of the same material.

In addition to the studies that carried the same involvement load indexes as the present study, studies including tasks with the same *need* and *search* indexes as the present study but different *evaluation* indexes (i.e., +, -, + = 2 vs. +, -, ++ = 3; Bao, 2015; Eckerth & Tavakoli, 2012; Karalık & Merç, 2016; Martínez-Fernández, 2008; Yang, Shintani, Li, & Zhang, 2017; Zou, 2017) were also further considered. The comparisons with such involvement load indexes included in these studies were manipulations of different degrees of *evaluation*, not the absence and presence of it, which was the case in the present study. There were cases of support for the ILH (e.g., Eckerth & Tavakoli, 2012; Yang, Shintani, Li, & Zhang, 2017; Zou, 2017) and also no support (e.g., Martínez-Fernández, 2008; Yang, Shintani, Li, & Zhang, 2017).

Due to the issues which make it difficult to make direct comparisons between the present study and the ones reviewed above, it seems that this study could be better discussed in relation to itself. The results will be discussed in relation to (1) task properties and (2) testing instruments.

5.1 Task properties

5.1.1 Word replacement: Participant performance

Word replacement was at the core of all the tasks/tests used in the study, because it was the only task type that rendered the two tasks different from each other in terms of the ILH. If there was a possible difference in task effectiveness in this study, it had to come from the fact that these tasks required word replacement in the opposite directions. It would have been the best option to have all the participants replace all the words correctly before they proceeded to the following tests. Indeed, the participants' job was made easier with the mini-dictionary provided for word meanings. In addition to equipping the participants with word meaning information, extra words were included in the wordlists to make every word replacement a conscious choice. Full scores from this test was very important to uncover the full workings of the two tasks. It was believed that only after eliciting full scores in the word replacement task would one consider that a possible difference in group performances could be attributed to the difference between the two tasks. In other words, fully correct responses for the word replacement task was considered as a prerequisite for assessing the effects of task type. However, this was not achieved. Among the failures, the most striking cases included two target words whose meanings are close to each other: *bungle* and *desecrate*. There were 19 cases of wrong pairings of *bungle-pollute* and 25 cases of *ruin-desecrate*. Considering that the participants were advanced level learners, such numbers of wrong replacements seemed to be excessively high.

This task was not designed as a test of vocabulary knowledge, but rather as a task for vocabulary learning. Establishing initial form-meaning connections for all target words, which would have been an important step towards their retention, was

not possible for all participants. Therefore, in the absence of full scores on this task by all participants, it becomes really difficult to argue that the target words were learned, let alone claiming that the different involvement load indexes of these tasks affected the learning of the target words differently.

5.1.2 Word replacement: Task design

A possible argument as to why the tasks failed to yield different effects could be related to their history of revision during the course of the study. These tasks had originally been designed as pen and paper tasks. As such the word replacement task would involve crossing out ten words in the text and writing their counterparts in their place. However, when the tasks were changed from pen and paper version to the electronic version for the sake of reaching the participants more easily during/after the pandemic, this dimension was lost. With the GoogleForms electronic versions, the word replacement task did not go far from an isolated matching exercise where the participants simply clicked on their choices in a drop-down menu (see Figure 30). This way, the word replacement task was not done on the text and the word pairs to be replaced with each other were far from the text. The loss of handwriting dimension in the task design might have reduced the intensity of the close engagement with the task words, target and otherwise, and the text. Such a change might have affected the target words more adversely than their high-frequency counterparts because, after all, the target words needed every bit of attention they could get while the counterparts were considered to be known anyway.

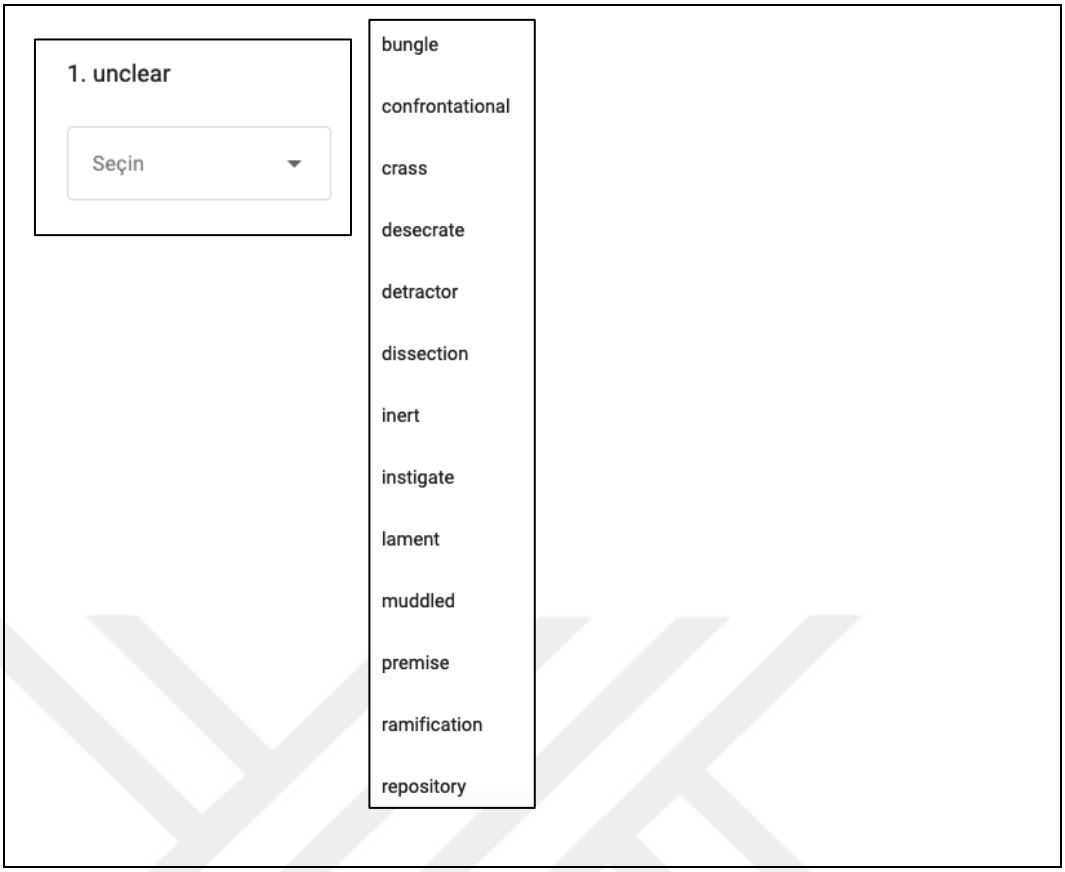

Figure 30. The electronic word replacement task sample
Note. *Seçin* means *Choose*.

In a study comparing three writing modalities, namely handwriting, writing with a touch keyboard, and writing with a regular keyboard, Mangen, Anda, Oxborough, and Brønnick (2015) found significantly better free recall of words for the handwriting modality than the keyboard modalities. The authors referred to the differences in terms of sensorimotor/graphomotor processes in handwriting and writing with a keyboard conditions:

> In handwriting, the writer has to graphomotorically form each letter from scratch – i.e, produce a graphic shape resembling as much as possible the standard shape of the specific letter. The graphomotor processes in the handwriting condition in our experiment may have facilitated a richer encoding of the words into long-term memory, resulting in better retrieval as evidenced in the free recall measure. (p. 240)

The handwriting dimension in the present study was not lost to keyboard typing for the reading+replacing tasks. Indeed, no way of word construction on the part of the

participants was possible because complete words were provided as candidates for correct counterparts. As such, possible cognitive benefits that can be associated with handwriting, or even any type of typing, were not relevant for the target words of the present study.

For the print versions used in the final round of data collection, it was not possible to use the original versions of the tasks, which required hand writing on the texts. This was not possible because those tasks inherently involved different processes and were not comparable with the GoogleForms version. Therefore, to maintain uniformity in the best way possible across the data collection sessions, the electronic formats were printed out for pen and paper sessions. Figure 31 shows how the word replacement task looked in the electronic-to-print versions.

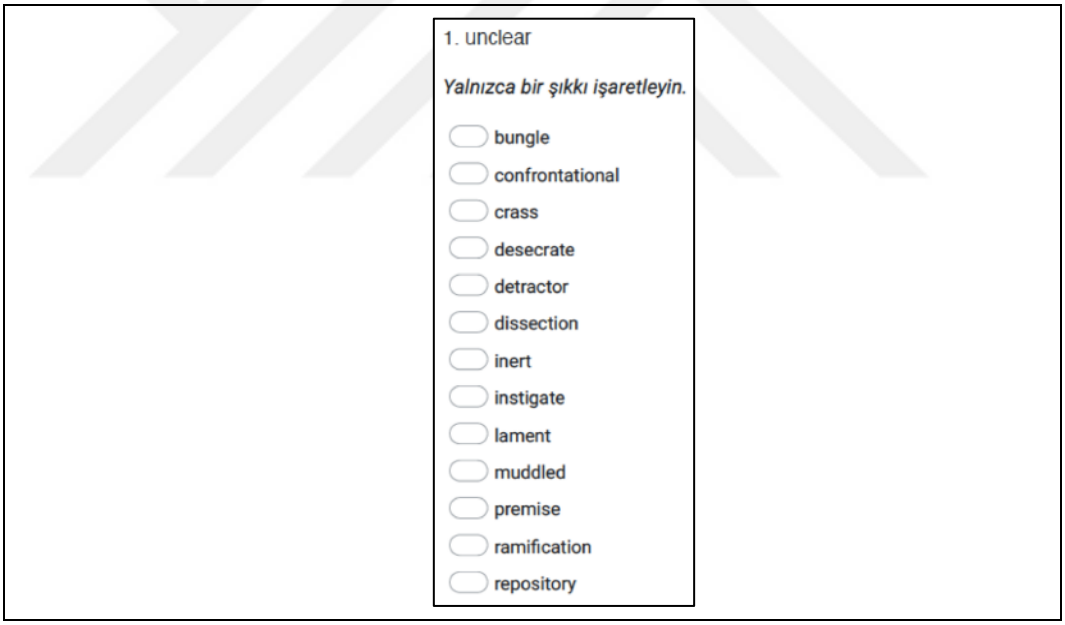

Figure 31. The electronic-to-print word replacement task sample
Note. Yalnızca bir şıkkı işaretleyin. means *Choose one option only.*

5.2 Testing instruments

Normally, an increase or no change should be observed in participants' word knowledge from a pretest to a posttest. In this study, however, there were cases which failed to confirm this. Two possible reasons seemed to explain why the target

words for which the participants claimed preknowledge sometimes became unknown later: (1) inconsistent judgments made for word knowledge and (2) multiple meanings of words.

To start with lack of consistency, there were 65 cases where the participants contradicted with their claim to know a target word, based on their self-reports. One of the participants, namely participant #1, who reported previous knowledge of six of the target words but who ended up getting 2 points (i.e., correct word meaning given for two words) in the immediate posttest, seemed to be a good candidate for further analysis (see Table 14).

Table 14. Participant #1's Pre- to Posttest Target Word Knowledge

Words preknowledge claimed for	Status in the immediate posttest		
		Response	Rater 1 and 2 initial[a] judgement
bungle	correct		
confrontational	wrong	"to face something"	2[b] - 2
detractor	wrong	"ruiner, slayer, someone who destroys something"	3[c] - 2
dissection		VKS option 2[d] selected	No judgement needed
muddled		VKS option 2 selected	No judgement needed
premise	correct		

[a]before resolution, [b]wrong, [c]correct synonym/translation given, [d]"*I have seen this word before, but I don't know what it means*."

This participant seemed to contradict with their previous statement rather immensely for two of the words (i.e., *dissection* and *muddled*) by choosing the "*I have seen this word before, but I don't know what it means*" option in the VKS. For the other two, this contradiction was less striking, with one of the words (i.e.,

detractor) even getting a positive rating from one of the raters initially. One thing that might deserve a thought is what participants understand from knowing a word. When they claim that they know a word, could it also be that they actually agree to *having seen that word before*, as in the VKS option 2?

Another point of consideration could be the use of self-report as a measure to assess preknowledge. Because the participants were not asked to indicate their knowledge of the words that they claimed to know, only their claims were taken as reference in scoring. The discrepancies of this kind, which were found between self-reports (as in the preknowledge test) and in practice (as in the posttests), could be caused by many factors. Recently, for the field of psychology, Dang, King, and Inzlicht (2020) pointed out how "across a series of domains, recent meta-analyses and large-scale investigations have consistently found that self-report and behavioral measures of the same construct were weakly correlated" (p. 267). One compelling argument they provided in their discussion concerned the different things behavioral measures and self-report measures likely *tap*. It was argued that the former focuses on the best performance of a given person whereas the latter focuses on usual performance. Overall, 28 participants showed a decrease in their knowledge of the target words from the preknowledge test to the immediate posttest, two tests which were given in the same posttest session. Additionally, the number of participants who indicated previous knowledge of a target word, but who later indicated a lack of knowledge in either of the posttests (i.e. by choosing the VKS scale categories I or II) can be seen in Figure 32. This analysis came from the participants' self-reports directly, without any assessment done on them. Therefore, it worked more as a reflection of how consistent the participants were in their claims to know the target

words. It should also be noted that some participants both marked category II and provided word meaning.

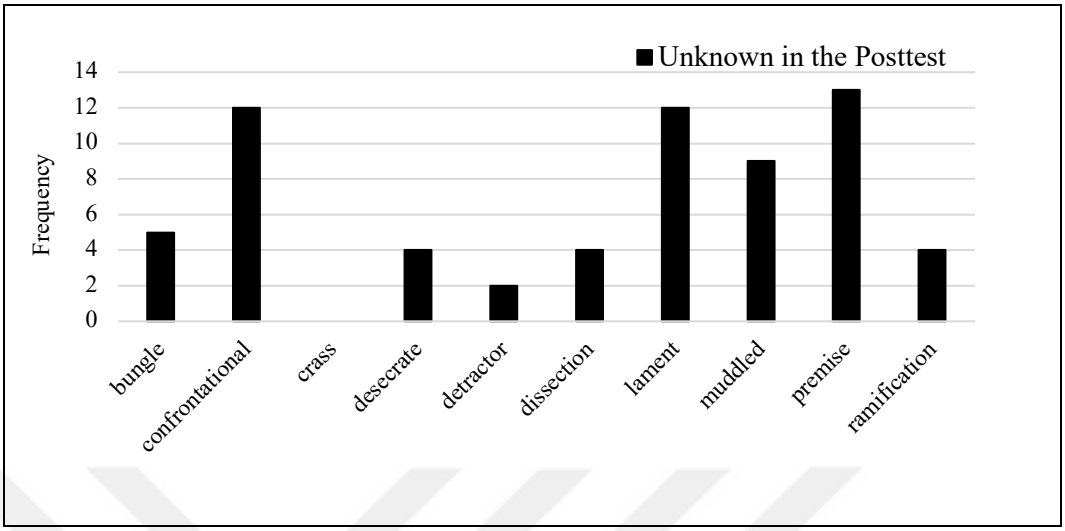

Figure 32. Target words marked 'known' initially but 'unknown' later

The second possible reason for an allegedly known word to become unknown in the posttest could be the specific word meaning targeted as part of the study. As is known, words can have multiple meanings. Some words can be used in their literal sense, metaphorically, or with one of the many meanings they have. This, indeed, is a matter of the *depth* of word knowledge, which Milton (2009) explains as follows:

> Depth is generally used to refer to a wide variety of word characteristics, including the shades of meaning a word may carry, its connotations and collocations, the phrases and patterns of use it is likely to be found in, and the associations the word creates in the mind of the user. (p. 149).

It was later realized that some of the word meanings provided by the participants in the posttests were not incorrect, but simply not the ones this study aimed to teach. Participant #31's responses serve an example of this (see Table 15).

Table 15. Participant #31's Pre- to Posttest Target Word Knowledge

Words preknowledge claimed for	Status in the immediate posttest		
		Response	Rater 1 and 2 initial[a] judgement
confrontational	correct		
dissection	wrong	"cutting corpses open for analysis"	2 - 2
lament	wrong	VKS option 2[b] selected	No judgement needed
premise	wrong	VKS option 2 selected	No judgement needed

[a]before resolution, [b]"*I have seen this word before, but I don't know what it means.*"

As can be seen, Participant #31 defined the word *dissection* in its classical sense, or with its more concrete meaning, if not classical. Therefore, this participant's claim to know *dissection* can be considered absolutely valid. If they had provided the targeted meaning of this word, which was provided in the mini-dictionary as "*dissect (verb): to examine something carefully in order to understand it*", they would have retained their score for this word. It is highly likely that this participant will understand this word upon encountering it in its task meaning elsewhere. However, word knowledge in this study was constrained to specific meanings.

Table 16 shows participant responses where the meaning provided was correct, but not targeted in the present study and thus considered unknown. Of the 11 different participants who lost a score for *lament*, 7 had indicated preknowledge. Of the 11 different participants who lost a score for *dissection*, 9 had indicated preknowledge. This was one of the reasons why the knowledge of the target words decreased from pre- to posttests.

Table 16. Participant Responses Treated as Incorrect for Words with Multiple Meanings

Target Words	Posttest Meanings Provided	Number of Cases
lament	to sob, extremely sadden about something to be sad about s[o]m[e]th[ing] (to) grieve (x2) ağıt[a] (x2) grief from someone's back mourn (x2) be upset about to grieve or w[]eep woe to remember someone who passed away	13
dissection	bölme[b] to dismantle a thing into its parts cutting corpses open for analysis split somet[h]ing into pieces tearing apart ayırma[b] Like sep[a]rating something cut up slice to divide things into pieces cutting something into smaller pieces cutting apart	12

[a]meaning intended = a sad song. [b]meaning intended = to divide.

In hindsight, though, target words with multiple meanings could have been excluded. This would have helped getting a more precise understanding of target word gains or lack thereof. Indeed, with the problems inherent in self-report preknowledge tests, it would have been even better to ask the participants to define the words that they claimed to know. This would have made the preknowledge test used in the present study better (including its words with multiple meanings), especially if the participants had also been asked to define the words with all their possible meanings. Still, the ideal scenario would be to use words with a single meaning and ask participants to define them, instead of just claiming to know them or not. Such inconsistencies, unfortunately, made many discussions that could have been built upon a healthy link between a pretest and a posttest untenable.

From a vocabulary knowledge point of view, it is also important to question why some of the participants defined certain words with meanings which were not targeted in the present study. It is hard to tell whether these participants learned the task meanings for these words or not. However, what is known for sure is that the task meanings were not their automatic/first choice for definition. One possible explanation could be they did not pay attention to these new meanings or they needed more time to add the new meanings to their repertoire. As Nation (2020) suggested, word knowledge "involves knowledge of a variety of different aspects of knowledge, and these aspects of knowledge can be known to different levels of strength and detail, and to different levels of fluency" (p. 15). Therefore, it is important to keep in mind that the relationship between word form and word meaning is rarely unidimensional.

In addition to participant behavior, the type of posttest used in this study needs further consideration. In their meta-analysis, Yanagisawa and Webb (2021) discussed the following in relation to delayed posttests, learning, involvement load, and test types:

> At the delayed posttest, learning gains tend to show decay due to the passage of time, and differences in gains across tasks with different [involvement loads] may be revealed more clearly with more sensitive tests. Form recognition and form-meaning recognition tests are sensitive to smaller degrees of knowledge and may capture differences in gains that less sensitive tests such as form-meaning recall and use tests cannot capture at the delayed posttest. In contrast, when learning gains are measured immediately after learning with recognition tests, learners may easily recognize target words they were exposed to even during a low-[involvement load] task. (p. 520)

With the VKS, this study used tests of meaning recall, though lower-level word knowledge (i.e., score categories I and II) was also recognized. It could have been better to use a form/ meaning recognition test as the immediate posttest. Although

this could have made the participants' task easier, as Yanagisawa and Webb argues, it could have given them an additional exposure to the low-frequency target words.

CHAPTER 6

CONCLUSION

In this study, completing tasks with different involvement load indexes did not lead to a difference in the early and long-term retention of low-frequency words. The participants' early and long-term knowledge of the target words was found to be lower than the preknowledge they had indicated. It was the long-term knowledge of the target words which was significantly lower than that of preknowledge. Multiple word meanings and certain inconsistencies in participant self-reports seemed to be part of the reasons why the preknowledge of the target words was always higher than retention.

Overall, the trio *muddled, premise*, and *confrontational*, the target words with the best preknowledge rates, tended to achieve the best retention rates in this study, whereas *lament* and *bungle* the worst. The trio *bungle*, *crass*, and *detractor,* the target words with the lowest preknowledge rates, most commonly (1) remained familiar but unknown and this tendency was followed by (2) gained familiarity over time (with meaning unknown). The best common scenario for *bungle*, *crass*, and *detractor*, therefore, seemed just familiarity.

When the participants are considered as a whole, the learning that took place seemed quite low. It turned out that 49.1% of the participants showed knowledge of 3 words or less in the short term. In the long term, 60.4% of the participants showed such knowledge. The two most common patterns of target word retention were in the form of maintaining familiarity or gaining familiarity with the target word, without learning word meaning.

6.1 Pedagogical implications

For the learning of individual low-frequency words and/or one of the meanings they have, language exercises or learner habits which prioritize the number of encounters can be given more emphasis in the L2 classroom. Instead of engaging with the target words once in contexts which require different degrees of learner involvement, learners can be encouraged to maximize their exposure to target words. This quantity approach could be strengthened with the quality of the exercises. In other words, tasks with different degrees of involvement can be visited multiple times. An advanced learner who has just learned a new low-frequency word, as a result of engaging in a language exercise or as a by-product of an effort to make their text lexically more sophisticated, can put further individual effort into seeing the new word again, in ways which they think best fit their learning style.

Another implication concerns material choice/adaptation for the L2 classroom. This study used an authentic text; however, some of the target words were very close to each other. A more balanced distribution of target words can be achieved in order for learners to work with each unknown word more independently. In other words, authenticity of the reading text and unknown word location can be given equal emphasis. If this is not possible with the authentic material, it can be adapted by making changes to text length or text content.

Use of handwriting in the language classroom, particularly for vocabulary learning, could be an issue worthy of consideration for language learners. Being exposed to unknown word forms in electronic tasks may prevent learners from giving their full attention to these words or give them a false sense of mastery. Therefore, writing words on paper, and other ways of closer engagement with them, could be encouraged in the L2 classroom.

Finally, the range of topics selected for reading texts could be varied. In line with the above idea of maximizing the quantity of encounters with the unknown words, learners can be exposed to such words in texts of differing topics, if possible. This way, learners' willingness to engage in the meaning-making process can be kept alive.

6.2 Limitations and suggestions for further research

Although this study aimed to investigate advanced learners' retention of low-frequency vocabulary in relation to the ILH from a different task point of view, the results were limited to the materials and methods used, many of which were discussed in the previous sections. The low sample size was a limitation. Additionally, the original tasks, which had been designed for in-person implementation and thus believed to encourage different cognitive and implementational processes, could not be used. Another limitation was the reliance on participants' self-report knowledge only in the preknowledge test. As for the data collection procedure, in order to reach an adequate number of participants, the data collection process of this study spanned over a considerable length of time. During that time, due to the changes necessitated by the pandemic and for the sake of convenience, participation in the study took different forms: online and face-to-face. The online sessions were naturally far less controlled than face-to-face sessions. The conditions under which the participants completed the tasks were not known well. The quality of the technological tools used, the internet connection, and the environment were all variables which could not be controlled but could have had an impact on participants' performance. Similarly, during the sessions held for the face-to-face data collection process, a case of compulsory transfer to another room and

outside noise might have distracted some of the participants. Finally, in the delayed posttest, the participants were warned to do the tasks in the order they were presented, because the form recall test had to precede the VKS. However, in both of the participation conditions, there was the risk of participants not acting accordingly or navigating between the tasks.

Future research could retest such task pairs as used in the present study by preserving the real-life idea of text revision throughout the task and using a larger sample size. An adequate sample size which includes only the participants with complete success in the word replacement task and no previous knowledge of the target words might show the effectiveness, or lack thereof, of the tasks more clearly. In terms of materials, checking preknowledge at the time of posttest might have been confusing for the participants. They might have found it difficult to think back to their previous state of word knowledge. Therefore, the preknowledge test can be given before the tasks proper, and it should require the demonstration of knowledge claimed for a given target word. As for the posttests, if the target words used in the study have multiple meanings, the posttest should ask all possible word meanings because participants could provide a meaning which is not incorrect but irrelevant for the task at hand. Still, a better solution would be to use a recognition test, rather than a recall test. This would ensure a participant response which will show whether learning has occurred or not, provided that the answer is a conscious choice.

APPENDIX A

TASK A (LOWTEXT)

AKTİVİTE *(Activity)*

Açıklamalar:

- Size bir okuma metni (Text) verilecek.
- Sizden metni okumanız ve metinde **kalın** harflerle yazılmış her bir kelimeyi, kelime listesindeki (Word List) **eş anlamlısı** ile değiştirmeniz beklenmektedir.
- Bilmediğiniz kelimelerin anlamına size verilen küçük bir sözlük olan Mini-Dictionary aracılığıyla bakınız.
- Her bir kelimeyi sadece bir defa kullanınız.
- Kelime listesinde (Word List) kullanmanıza gerek olmayan üç kelime bulunmaktadır.

(Instructions:

** You will be given a reading text (Text).*
** You are expected to read the text and replace each **boldfaced** word in the text with its **synonym** in the word list (Word List).*
** Please look up the words that you do not know in the Mini-Dictionary provided for you.*
** Please use each word only once.*
** In the word list (Word List), there are three words which you do not need to use.)*

Text	Word List

Does Influencer Grammar Matter?

In Southeast Asia, watchdog accounts call out misspelled and otherwise **unclear** English-language captions.

Cindy Cendana, an Indonesian beauty influencer, just wanted to know if any of her followers had been to the Japanese city of Himeji. So she posted the question — "Hand's up, who's been to #Himeji, Japan?" — on Instagram, along with a photo of herself in a blue dress.

Ms. Cendana could have anticipated a number of comments — restaurant recommendations, tips about visiting the feudal castle there — but certainly not this one: "If you put an apostrophe after 'hand,' it either means 'hand is' or something that belongs to the hand. In this case, it should be 'hands' as you're referring to multiple hands." (The post no longer appears in Ms. Cendana's feed.)

She wrote in a text message that she still remembers when she first saw the comment in February: "I was surprised, but happy at the same time because I had my clumsiness corrected for me."

The comment came from the anonymous Instagram account @englishbusters, which criticizes Indonesian influencers who **ruin** English grammar on social media. Its provocative posts — which can be in the form of a forensic, **foolish** **analysis** of a caption like "wanna coffee?" or "Why choosing between yoga and fashion when you can have them both?" — have divided Indonesian social media and press. Its followers appreciate its direct, **aggressive** tone; however, a **hater** will **complain** that it uses snark, and feel that its very **argument** borders on public shaming.

bungle

confrontational

crass

desecrate

detractor

dissection

inert

instigate

lament

muddled

premise

ramification

repository

@englishbusters follows a broader trend of calling out influencers for their antics, whether it's aimed to **pollute** American public lands or display general social entitlement. The account also has a lexicological precedent: It opened up shop a couple of years after a notorious Facebook page called "Singaporean Influencers and Bloggers Write ____ English and are Annoying AF" lit up Singapore, Indonesia's neighboring country.

Just as on other continents, the business of influence has flourished in Asia. The social media marketing company Socialbakers stated in a 2019 report that Instagram #sponcon in Asia had increased by 189 percent since last year. But does anyone really care whether those posts include spelling and grammatical errors?

Dennis Toh, a marketing and design lecturer at the Management Development Institute of Singapore and a founder of the Influencer Network, an agency with offices in Singapore, argues that the language does indeed matter. "When you are linguistically strong, that is an advantage," he said. "To put a point across, good grammar is a basic requirement."

The Influencer Network, Mr. Toh said, does not strictly review the syntax of influencer posts upon receiving briefs from brands; most of them write well anyway. But Okto Rohyadi, the account manager of the Jakarta-based marketing company Glitzmedia, said that grammar doesn't matter as much as voice, be it in English or Indonesian. "It does come down to how the influencers usually talk," he said.

Nelly Martin-Anatias, a lecturer at the School of Language and Culture, Auckland University of Technology, said that Indonesians still treat English as a foreign language, instead of a second language. "There is a tendency for Indonesians in big cities to code-switch between Indonesian and English," she said. "There is an ideology baked in the English language that it's the cool language, the successful language."

Why is English common among Indonesian influencers, though? "Language is the primary marked code to show or manipulate our identities. That's why English is common in the marketing world here," Ms. Martin-Anatias said.

Ms. Cendana, when asked whether she feels responsible to educate her followers, gave a short answer: "Yes."

The long answer?

"Much like actors, politicians, entrepreneurs, what they say should have a positive **effect**. Everyone is capable of influencing everybody. Gone are the days you put someone on a pedestal," Mr. Toh said.

KELİME DEĞİŞTİRME AKTİVİTESİ *(Word Replacement Activity)*

Aşağıda verilmiş olan, metinde geçen her bir kelime (1-10) için ★★ kelime listesinden (Word List) doğru eş anlamlı kelimeyi seçiniz.

1. unclear

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

2. ruin

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

3. foolish

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

4. analysis

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

5. aggressive

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

6. hater

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

7. complain

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

8. argument

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

9. pollute

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

10. effect

Yalnızca bir şıkkı işaretleyin.

- bungle
- confrontational
- crass
- desecrate
- detractor
- dissection
- inert
- instigate
- lament
- muddled
- premise
- ramification
- repository

★*(For each word below (1-10) which appeared in the text, please choose the correct synonym in the word list (Word List).)*

★★*(Choose one option only.)*

APPENDIX B

TASK B (HIGHTEXT)

AKTİVİTE *(Activity)*

Açıklamalar:

- Size bir okuma metni (Text) verilecek.
- Sizden metni okumanız ve metinde **kalın** harflerle yazılmış her bir kelimeyi, kelime listesindeki (Word List) **eş anlamlısı** ile değiştirmeniz beklenmektedir.
- Bilmediğiniz kelimelerin anlamına size verilen küçük bir sözlük olan Mini-Dictionary aracılığıyla bakınız.
- Her bir kelimeyi sadece bir defa kullanınız.
- Kelime listesinde (Word List) kullanmanıza gerek olmayan üç kelime bulunmaktadır.

(Instructions:

** You will be given a reading text (Text).*
** You are expected to read the text and replace each **boldfaced** word in the text with its **synonym** in the word list (Word List).*
** Please look up the words that you do not know in the Mini-Dictionary provided for you.*
** Please use each word only once.*
** In the word list (Word List), there are three words which you do not need to use.)*

Text

Does Influencer Grammar Matter?

In Southeast Asia, watchdog accounts call out misspelled and otherwise **muddled** English-language captions.

Cindy Cendana, an Indonesian beauty influencer, just wanted to know if any of her followers had been to the Japanese city of Himeji. So she posted the question — "Hand's up, who's been to #Himeji, Japan?" — on Instagram, along with a photo of herself in a blue dress.

Ms. Cendana could have anticipated a number of comments — restaurant recommendations, tips about visiting the feudal castle there — but certainly not this one: "If you put an apostrophe after 'hand,' it either means 'hand is' or something that belongs to the hand. In this case, it should be 'hands' as you're referring to multiple hands." (The post no longer appears in Ms. Cendana's feed.)

She wrote in a text message that she still remembers when she first saw the comment in February: "I was surprised, but happy at the same time because I had my clumsiness corrected for me."

The comment came from the anonymous Instagram account @englishbusters, which criticizes Indonesian influencers who **bungle** English grammar on social media. Its provocative posts — which can be in the form of a forensic, **crass dissection** of a caption like "wanna coffee?" or "Why choosing between yoga and fashion when you can have them both?" — have divided Indonesian social media and press. Its followers appreciate its direct, **confrontational** tone; however, a **detractor** will **lament** that it uses snark, and feel that its very **premise** borders on public shaming.

Word List

aggressive
analysis
argument
begin
complain
container
effect
foolish
hater
inactive
pollute
ruin
unclear

@englishbusters follows a broader trend of calling out influencers for their antics, whether it's aimed to **desecrate** American public lands or display general social entitlement. The account also has a lexicological precedent: It opened up shop a couple of years after a notorious Facebook page called "Singaporean Influencers and Bloggers Write ____ English and are Annoying AF" lit up Singapore, Indonesia's neighboring country.

Just as on other continents, the business of influence has flourished in Asia. The social media marketing company Socialbakers stated in a 2019 report that Instagram #sponcon in Asia had increased by 189 percent since last year. But does anyone really care whether those posts include spelling and grammatical errors?

Dennis Toh, a marketing and design lecturer at the Management Development Institute of Singapore and a founder of the Influencer Network, an agency with offices in Singapore, argues that the language does indeed matter. "When you are linguistically strong, that is an advantage," he said. "To put a point across, good grammar is a basic requirement."

The Influencer Network, Mr. Toh said, does not strictly review the syntax of influencer posts upon receiving briefs from brands; most of them write well anyway. But Okto Rohyadi, the account manager of the Jakarta-based marketing company Glitzmedia, said that grammar doesn't matter as much as voice, be it in English or Indonesian. "It does come down to how the influencers usually talk," he said.

Nelly Martin-Anatias, a lecturer at the School of Language and Culture, Auckland University of Technology, said that Indonesians still treat English as a foreign language, instead of a second language. "There is a tendency for Indonesians in big cities to code-switch between Indonesian and English," she said. "There is an ideology baked in the English language that it's the cool language, the successful language."

Why is English common among Indonesian influencers, though? "Language is the primary marked code to show or manipulate our identities. That's why English is common in the marketing world here," Ms. Martin-Anatias said.

Ms. Cendana, when asked whether she feels responsible to educate her followers, gave a short answer: "Yes."

The long answer?

"Much like actors, politicians, entrepreneurs, what they say should have a positive **ramification**. Everyone is capable of influencing everybody. Gone are the days you put someone on a pedestal," Mr. Toh said.

KELİME DEĞİŞTİRME AKTİVİTES *(Word Replacement Activity)*

Aşağıda verilmiş olan, metinde geçen her bir kelime (1-10) için ★ kelime listesinden (Word List) doğru eş anlamlı kelimeyi seçiniz.

1. muddled

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

★★

2. bungle

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

3. crass

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

4. dissection

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

5. confrontational

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

6. detractor

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

7. lament

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

8. premise

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

9. desecrate

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

10. ramification

Yalnızca bir şıkkı işaretleyin.

- aggressive
- analysis
- argument
- begin
- complain
- container
- effect
- foolish
- hater
- inactive
- pollute
- ruin
- unclear

★*(For each word below (1-10) which appeared in the text, please choose the correct synonym in the word list (Word List).)*

★★*(Choose one option only.)*

APPENDIX C

THE MINI-DICTIONARY

MINI-DICTIONARY

bungle	(verb) to fail to do something properly, because you have made stupid mistakes – used especially in news reports
confrontational	(adjective) likely to cause arguments or make people angry
crass	(adjective) behaving in a stupid and offensive way which shows that you do not understand or care about other people's feelings
desecrate	(verb) to spoil or damage something holy or respected
detractor	(noun) someone who says bad things about someone or something, in order to make them seem less good than they really are
dissection	**dissect** (verb) to examine something carefully in order to understand it
inert	(adjective) not willing to do anything
instigate	(verb) to make a process start, especially one relating to law or politics
lament	(verb) to express annoyance or disappointment about something you think is unsatisfactory or unfair
muddled	(adjective) confused
premise	(noun) a statement or idea that you accept as true and use as a base for developing other ideas
ramification	(noun) an additional result of something you do, which may not have been clear when you first decided to do it
repository	(noun) a place or container in which large quantities of something are stored

APPENDIX D

READING COMPREHENSION QUESTIONS

OKUDUĞUNU ANLAMA AKTİVİTESİ ★

Verilen ifadeler için (1-5) doğru cevabı (a, b, c ya da d) seçiniz. ★★

1. What was unusual about the comment Ms. Cendana received in response to her question?

- a. The comment listed the worst places to eat.
- b. The comment was about top tourist attractions.
- c. The comment focused on her grammar.
- d. The comment was an insult to her appearance.

2. Which of the following does NOT describe Ms. Cendana's feelings after reading the comment?

- a. content
- b. proud
- c. satisfied
- d. grateful

3. What kind of reaction has @englishbusters produced?

- a. It has gained support but also faced opposition.
- b. Both Indonesian social media and press are negative about its posts.
- c. It faced resistance in relation to reports and news.
- d. The public feels embarrassed because of its posts.

4. Which one of the below is NOT true according to the text?

- ◯ a. Challenging influencers is already a common situation.
- ◯ b. The role of influence on the market is acknowledged only in Asia.
- ◯ c. Indonesians in different cities may use Indonesian differently.
- ◯ d. Ms. Cendana cares about affecting her followers positively.

5. It can be inferred from Dennis Toh and Okto Rohyadi's comments that ...

- ◯ a. use of grammar is the most important criterion for marketing companies.
- ◯ b. companies should check influencers' posts for grammatical errors.
- ◯ c. different views exist about the importance of grammar in influencers' posts.
- ◯ d. both English and Indonesian are great languages for communication.

★*(Reading Comprehension Activity)*

★★*(Please choose the correct answer (a, b, c, or d) for the statements given (1-5).*

APPENDIX E

THE SURVEY OF TASK INTERACTIVENESS

KENDİNİ DEĞERLENDİRME AKTİVİTESİ★

Aşağıda verilen her bir ifade için katılma derecenizi gösteren numarayı seçiniz. ★★

1. The reading text was suitable for my English level.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

2. I was familiar with the topic of the reading text before participating in this study.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

3. I enjoyed reading the text.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. I enjoyed completing the tests that followed it.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. I was familiar with the task (replacing words) before participating in this study.

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

★*(Self-Evaluation Activity)*

★★*(For each statement below, please choose the number which shows your agreement.)*

APPENDIX F

THE TEST OF PREKNOWLEDGE

Açıklamalar *(Instructions:)*

Lütfen aşağıdaki kelimeleri **bugünkü oturumdan önce** bilip bilmediğinizi bildiriniz.

*(Please indicate whether you knew the words below **before today's session**.)*

	I knew this word before our meeting.	I didn't know this word before our meeting.
bungle	○	○
confrontational	○	○
crass	○	○
desecrate	○	○
detractor	○	○
dissection	○	○
lament	○	○
muddled	○	○
premise	○	○
ramification	○	○

APPENDIX G

IMMEDIATE AND DELAYED POSTTESTS: THE VKS

NOTE. THE INSTRUCTION BELOW APPEARED ONLY IN THE IMMEDIATE POSTTEST

Bu kısım için, lütfen bilginizi GENEL OLARAK göz önünde bulundurun.
Sadece bugünkü oturumun öncesi olarak DÜŞÜNMEYİN.

(For this section, please consider your OVERALL knowledge. Do NOT consider your previous knowledge only.)

Açıklamalar: *(Instructions:)*

Lütfen aşağıdaki her bir setin başında verilen kelime hakkındaki bilginizi **seçim yaparak (I veya II için)** ya da **yazarak (III, IV veya IV+V için)** gösteriniz.

*(For the word given at the top of each set below, please indicate your knowledge by **choosing (for I or II)** or **writing (for III, IV or IV+V).**)*

1. ramification

○ I I don't remember having seen this word before.
○ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

2. bungle

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

3. muddled

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

4. confrontational

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

5. lament

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

6. crass

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

7. premise

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

8. desecrate

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

9. dissection

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

10. detractor

◯ I I don't remember having seen this word before.
◯ II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I think it means _____. (synonym or translation)
IV I know this word. It means _____. (synonym or translation)
V I can use this word in a sentence: _____. (Write a sentence.) *(If you do this section, please also do Section IV.)*

APPENDIX H

THE TEST OF FORM RECALL

Açıklamalar: Lütfen cümleleri okuyunuz ve kalın yazılmış kelimeleri **eş anlamlılarıyla** değiştiriniz. Verilen kutuya sadece eş anlamlı kelimeyi yazınız.

*(Instructions: Please read the sentences and replace the boldfaced words with their **synonyms**. Write only the synonym in the box given.)*

1.

The issue is very serious, so the **effect** will be far beyond the loss of cash by more than a million families.

2.

There also seems to be broad opinion that chapters 4-14 are **unclear**, sometimes beyond intelligent interpretation.

3.

She talks of her **aggressive** attitude toward them during the early days of her captivity.

4.

Visitors to the island used to **complain** that the area did not have any cheap restaurants.

5.

The fundamental **argument** of the publication is that early design for space travel was influenced largely by science fiction.

6.

It seems **foolish** and inappropriate, but I understand his need to record this terrible scene.

7.

It was so sad to watch the players **ruin** five goal attempts by misdirecting shots that could have easily been placed into the nets.

8.

He believed that such irresponsible actions would **pollute** family values.

9.

This review will not be an extensive **analysis** of this film.

10.

After all, whether you are a fan or **hater**, you are anonymous on the internet.

APPENDIX I

ETHICS COMMITTEE APPROVAL

T.C. BOĞAZİÇİ ÜNİVERSİTESİ
Sosyal ve Beşeri Bilimler İnsan Araştırmaları Etik Kurulu (SBİNAREK)

11.03.2019

Dr. Öğr. Üyesi Şebnem Yalçın,
Boğaziçi Üniversitesi,
Eğitim Fakültesi,
Yabancı Diller Eğitimi Bölümü
34342 Bebek / İstanbul

Sayın Araştırmacı,

"Advanced Learners'Retention of Low-Frequency L2 Vocabulary: The Effects of Task Type and Interaction" başlıklı projeniz ile Boğaziçi Üniversitesi Sosyal ve Beşeri Bilimler İnsan Araştırmaları Etik Kurulu (SBİNAREK)'e yaptığınız 2019/05 kayıt numaralı başvuru 08.03.2019 tarihli ve 2019/03 sayılı kurul toplantısında incelenmiş ve projenize etik onay verilmesine karar verilmiştir.

Saygılarımızla, bilgilerinizi rica ederiz.

Doç. Dr. Osman Sabri Kıratlı (Başkan)
Uygulamalı Bilimler Yüksek Okulu
Uluslararası Ticaret Bölümü
Boğaziçi Üniversitesi, İstanbul

Prof. Dr. Fatoş Gökşen (Üye)
Fen Edebiyat Fakültesi
Sosyoloji Bölümü
Koç Üniversitesi, İstanbul

Dr. Öğr. Üyesi C. Taylan Acar (Üye)
Fen-Edebiyat Fakültesi
Sosyoloji Bölümü
Boğaziçi Üniversitesi, İstanbul

Dr. Öğr. Üyesi Işıl Erduyan (Üye)
Eğitim Fakültesi
Yabancı Diller Eğitimi Bölümü
Boğaziçi Üniversitesi, İstanbul

Dr. Öğr. Üyesi Selcan Kaynak (Üye)
İktisadi ve İdari Bilimler Fakültesi
Siyaset Bilimi ve Uluslararası İlişkiler Bölümü
Boğaziçi Üniversitesi, İstanbul

Dr. Öğr. Üyesi Nur Soylu Yalçınkaya
Fen-Edebiyat Fakültesi
Psikoloji Bölümü
Boğaziçi Üniversitesi, İstanbul

Öğr. Gör. Dr. Suzan Üsküdarlı (Üye)
Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi, İstanbul

1 / 1

REFERENCES

Arnaud, P. J. L., & Savignon, S. J. (1997). Rare words, complex lexical units and the advanced learner. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 157-173). Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

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

Bardel, C. (2016). The lexicon of advanced L2 learners. In K. Hyltenstam (Ed.), *Advanced proficiency and exceptional ability in second languages* (pp. 73-109). Boston/Berlin: De Gruyter Mouton.

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

Byrnes, H. (2004). Toward academic-level foreign language abilities: reconsidering foundational assumptions, expanding pedagogical options. In B. L. Leaver & B. Shekhtman (Eds.), *Developing professional-level language proficiency* (pp. 34-58). Cambridge: Cambridge University Press.

Byrnes, H. (2006). Locating the advanced learner in theory, research, and educational practice: an introduction. In H. Byrnes, H. Weger-Guntharp, & K. A. Sprang (Eds.), *Educating for advanced foreign language capacities: constructs, curriculum, instruction, assessment* (pp. 1-14). Washington, D.C.: Georgetown University Press.

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning, 71*(3), 834-871.

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.

Dang, J., King, K. M. , & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences, 24*(4), 267-269.

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

Eşdeğer Sınavlar ve Diğer Muafiyet Kuralları [*Tests Equivalent to BUEPT and other requirements for exemption*]. (n.d.). Retrieved from https://yadyok.boun.edu.tr/ tr/buyes/muafiyet

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

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273-293.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

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

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

Hu, H. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: involvement load hypothesis versus technique feature analysis. *System, 56*, 28-39.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal, 96*(4), 544- 557.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349-381). Malden, MA: Blackwell Publishing Ltd.

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539-558.

Jahangiri, K., & Abilipour, I. (2014). Effects of collaboration and exercise type on incidental vocabulary learning: evidence against involvement load hypothesis. *Procedia- Social and Behavioral Sciences, 98*, 704-712.

Karalık, T., & Merç. A (2016). The effect of task-induced involvement load on incidental vocabulary acquisition. *Mustafa Kemal University Journal of Graduate School of Social Sciences, 13*(35), 77-92.

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.

Kim, Y. (2008a). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285-325.

Kim, Y. (2008b). The contribution of collaborative and individual tasks to the acquisition of L2 vocabulary. *The Modern Language Journal, 92*(1), 114-130.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly, 50*(4), 976-987.

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

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75*(4), 440-448.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp.126-132). London: Macmillan.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: a rationale for pedagogy* (pp. 20-34). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59*(4), 567-587.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook, 5*, 223-250.

Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 351-368). London: Routledge.

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., & Rozovski-Roitblat, B. (2015). Retention of new words: quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research, 19*(6), 687-711.

Mangen, A., Anda, L. G., Oxborough, G. H., & Brønnick, K. (2015). Handwriting versus keyboard writing: Effect on word recall. *Journal of Writing Research, 7*(2), 227-247.

Martínez-Fernández, A. (2008). Revisiting the involvement load hypothesis: awareness, type of task and type of item. In M. Bowles, R. Foote, S. Perpiñán & R. Bhatt (Eds.), *Selected Proceedings of the 2007 Second Language Research Forum* (pp. 210-228). Somerville, MA: Cascadilla Proceedings Project.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Nassaji, H., & Hu, H. M. (2012). The relationship between task-induced involvement load and learning new words from context. *IRAL, 50*(1), 69-86.

Nation, I. S. P. (2001a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2001b). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. M. Koski, R. D. Sell, and B. Wårvik (Eds.), *Language, learning, literature: Studies presented to Håkan Ringbom (English Department Publications 4)* (pp. 167-181). Turku: Åbo Akademi University.

Nation, P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from https://www.wgtn .ac.nz/lals/about/staff/paul-nation

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1-16.

Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 15-29). London: Routledge.

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

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*(1), 9-29.

Paribakht, T. S., & Wesche, M. (1996). Enhancing vocabulary acquisition through reading: a hierarchy of text-related exercise types. *The Canadian Modern Language Review, 52*(2), 155-178.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: a rationale for pedagogy* (pp. 174-200). Cambridge: Cambridge University Press.

Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66-82.

Polio, C., Fleck, C., & Leder, N. (1998). "If I only had more time:" ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*(1), 43-68.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Sanz, C. (1997). Experimental tasks in SLA research: Amount of production, modality, memory, and production processes. In W. R. Glass & A. T. Pérez-Leroux (Eds.), *Contemporary perspectives on the acquisition of Spanish, Vol. 2: Production, processing and comprehension* (pp. 41-56). Somerville, MA: Cascadilla Press.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503.

Sinclair, J. McH., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160). London: Longman.

Swain, M., & Lapkin, S. (2002). Talking it through: two French immersion learners' response to reformulation. *International Journal of Educational Research, 37*(3-4), 285-304.

Thornbury, S. (1997). Reformulation and reconstruction: tasks that promote 'noticing'. *ELT Journal, 51(*4), 326-335.

Türkçede Batı Kökenli Kelimeler Sözlüğü [The Dictionary of the Words of Western Origin in Turkish] (n.d.). Retrieved from https://sozluk.gov.tr/

Understanding Your TOEFL iBT Scores. (n.d.). Retrieved from https://www.ets.org/toefl/test-takers/ibt/scores/understand-scores.html

Vilkaitė-Lozdienė, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81-96). London: Routledge.

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225-239). London: Routledge.

Widianto, S. (2019, September 18). Does influencer grammar matter? *The New York Times*. Retrieved from https://www.nytimes.com/

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning, 71*(2), 487-536.

Yang, Y., Shintani, N., Li, S., & Zhang, Y. (2017). The effectiveness of post-reading word-focused activities and their associations with working memory. *System, 70*, 38-49.

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: extending the evaluation component of the involvement load hypothesis. *Language Teaching Research, 21*(1), 54-75.