The Involvement Load Hypothesis and Its Impact on Vocabulary Learning

Mandana Hazrat

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Linguistics

School of Cultures, Languages, and Linguistics
The University of Auckland

2020

Abstract

The Involvement Load Hypothesis (ILH) proposed by Laufer and Hulstijn (2001) is a framework to operationalize elaborate processing for L2 vocabulary learning. The hypothesis assumes that vocabulary learning is conditional upon a motivational factor, need, and two cognitive factors, search and evaluation, in vocabulary-learning tasks. Need was given three degrees of minus (-), moderate (+), and strong (++) depending on whether words need to be learned at all (-) and whether the motivation for learning is externally imposed (+) or is self-imposed (++). Search has two degrees of minus (-) and plus (+) depending on whether the meanings of words are given in a task (-) or must be looked up in a dictionary (+). Evaluation was given three degrees of minus (-), moderate (+), and strong (++) depending on whether words are elaborated at all (-) and whether elaboration is limited to making comparisons between words (+) or is in the form of using words in original contexts (++). The sum of need, search, and evaluation factors with their degrees of prominence is called the task involvement load. The hypothesis assumes that tasks inducing a higher involvement load are more effective for vocabulary learning compared to tasks inducing a lower involvement load.

The strengths of the hypothesis are its potential instructional applicability and its simplicity. However, its weaknesses have been discussed in the literature, including uncertainty about the weight of the factors for vocabulary learning, uncertainty about the impact of distribution of the factors, and the limited range of scores the ILH gives to tasks. Considering this, the aim of the study was to develop the hypothesis while preserving its simplicity. To this end, ten groups of learners, including nine experimental groups and a control group at the intermediate level of proficiency, participated in the study. The treatment for each experimental group was designed to induce a specific combination of the factors. Before the treatment, the groups took a vocabulary pre-test and a homogeneity test. After the treatment, they took three vocabulary post-tests measuring active recall, passive recall, and passive recognition of words.

The results could extend the evaluation factor by indicating that evaluation can be given four degrees of prominence rather than three. This extension may contribute to differentiating between a wider range of activities in terms of their effectiveness for vocabulary learning. The findings could also re-define the need factor by suggesting that need may consist of two different types of motivation, namely self-driven and task-driven motivation for learning, rather than being a motivational factor with three degrees. Re-defining the need factor may enrich the hypothesis by providing a broader view of motivation. The results also clarified the role of

search by indicating that search does not have a fixed degree of prominence and its contribution to vocabulary learning depends on the type of evaluation it is combined with and the target aspect of vocabulary learning. It may provide a more precise estimate of the power of this factor in a variety of vocabulary-learning tasks. It was also concluded that, contrary to the ILH assumptions, task involvement indices are not good indicators of task effectiveness. Task effectiveness may depend on the type of evaluation, presence or absence of search, and the target aspect of vocabulary. The suggested modifications may make the hypothesis more accurate for designing vocabulary-learning tasks and judging their effectiveness.

Dedication

To Shahyad, my husband, and to my parents and siblings who supported me incredibly during my PhD programme

Acknowledgements

I would like to express my sincere gratitude to Professor John Read, Dr Louisa Buckingham, and Dr Rosemary Erlam who had been my first and second supervisors at different periods during my PhD. Their thoughtful comments, incredible support, and encouragement made my PhD the most enjoyable part of my life.

I would like to thank my examiners, Professor Stuart Webb and Associate Professor Jonathan Newton, who provided interesting and insightful comments.

I also would like to express my appreciation to all the students who eagerly participated in this study and to my friends and colleagues who facilitated data collection in one way or another:

Dr Morteza Bagheri Sangachin, Dr Hamideh Baggali Basmenj, Nasrin Yousefi, Yashar Abbas Zadeh, Melody Dooka, Arezoo Nikpoor, Hadi Chaiforooshan, Kiarash Kasaee, and Roya Mehri

Table of Contents

List of Tables

List of Figures

Acronyms and Symbols

AEF	American English File
BNC	British National Corpus
CEFR	Common European Framework
CF	The Consent Form
CLT	Communicative Language Teaching
COCA	Corpus of Contemporary American English
CW	Composition-writing
E	Evaluation factor
EFI	Form focused instruction
FFL	English as a Foreign Language
FonF	Focus on form
FonFs	Focus on forms
II	Involvement Index
ILH	Involvement Load Hypothesis
L1	First Language
L2	Second Language
N	Need factor
No.	Number
p	Probability value
P	Productive memory retrieval
PIS	Participants' Information Sheet
R (r)	Effect size
R	Receptive memory retrieval
S	Search factor
SW	Sentence-writing
TBLT	Task-based language teaching
TFA	Technique Feature Analysis
U	Mann-Whitney U
Z	z score
-	No involvement
+	Moderate degree of involvement
++	Strong degree of involvement

Chapter 1: Introduction

1.1 The Place of Vocabulary in Language Learning

The importance of vocabulary in learning a language cannot be ignored. As Milton (2009) states, among factors that contribute to language performance, vocabulary is "the biggest element of knowledge" (p. 179). Meara (2009, p. 34) mentions that language learners with a larger vocabulary communicate in a target language more effectively compared to learners with a more detailed command of a smaller vocabulary. The contribution of vocabulary knowledge to successful language performance has been supported by research (e.g., Barcroft, 2004 a; Milton, 2013; Schmitt, 2010 a; Stæhr, 2009). Evidence shows that vocabulary knowledge plays an essential role in listening comprehension (Stæhr, 2009; Vandergrift & Baker, 2015), reading comprehension (Schmitt, Jiang & Grabe, 2011; Wallace, 2008), and oral and written production (Koizumi, 2005; Nation, 2008).

From a learner's perspective, particular literacy tasks require a certain level of vocabulary knowledge. Nation (2006), by analysing the number of words in each level of frequency in several texts, established that to read a novel or newspaper, learners should know 8,000-9,000 words and to be familiar with most words in a children's movie, knowing over 7,000 words is necessary. Milton's (2009, p. 186) estimate of the number of words a learner needs to know is lower. Based on the relationship between learners' vocabulary size and their proficiency test scores, he states that English learners may need to know nearly 3000 words to progress from elementary to intermediate level of proficiency. While they progress toward the higher level of proficiency (advanced), they may gain wider vocabulary knowledge and reach 4000 words. According to Milton, a proficient learner of English has the knowledge of 5000 words. Another study by Stæhr (2009) provides a similar estimate. In this study, it is shown that a language learner should know 5000 word families for successful listening comprehension (at the advanced level). Later, Schmitt, Jiang, and Grabe's (2011) study resulted in numbers similar to Nation's estimate. It showed that learners need to know about 8000-9000 word families to read a variety of texts in English. Laufer (2013) also confirms this estimate by stating that learners need to know 8000 word families to read academic material independently. With knowledge of 5000 word families, they still need some assistance of a dictionary to manage the reading material.

Learning such a large number of words, although necessary for successful language performance, is a long-term process and may seem an unachievable goal for many language

learners. Schmitt (2010 a, p. 30) emphasizes the difficulty of vocabulary learning by stating that unlike grammar, which is a closed system with limited number of rules, vocabulary is an open-ended one and even for native speakers, vocabulary learning is a lifelong process. Schmitt (2010 a, p. 30) adds that the difficulty of vocabulary learning is greater for learners of English because English has a large vocabulary with loanwords from a large number of other languages and because English has semantically related words with no formal similarities. For English learners, this means learning a vast number of individual word forms.

In spite of the importance of vocabulary knowledge and the difficulty of vocabulary learning, vocabulary used to be disregarded and subordinated to other language elements such as grammar for a long time (Laufer & Nation, 2012, p. 163). However, later, the importance of expanding vocabulary knowledge and the difficulty of achieving this goal were recognized, which led to investigations about the effective ways of vocabulary learning. Views about vocabulary instruction have also undergone a paradigm shift that is explained in the following section.

1.2 Vocabulary Teaching and Learning

Formerly, more attention was given to grammar that, as a system with finite relations, is easier to handle compared with vocabulary in which relations are infinite (Carter, 1987; cited in Piquer Piriz, 2008, p. 220; Chacon-Beltrán, Abello-Contesse, & Torreblanca-Lopez, 2010, p. 1). The Grammar Translation method, with its emphasis on morphology and syntax, only viewed vocabulary as an aid to translation. In this method, no value was given to vocabulary learning and retrieval for communication. Later, the Audiolingual method emerged. The focus was still teaching morpho-syntactic features together with L2 sound system through memorization and drilling. The Audiolingual method was based on the behaviourist view of learning as habit formation and on poor beliefs about comprehension and cognitive processing. The substitution drills on which the method relies could hinder associating forms and meanings of words and could not promote vocabulary learning (Boers & Lindstromberg, 2008, p. 2).

Later, the introduction of Communicative Language Teaching (CLT) in 1980s shifted the focus of language learning to communicative functions (Boers & Lindstromberg, 2008, p. 3). In the early 1990s, research in the area of vocabulary increased (Schmitt, 2010 a, p. 28) and after a period of neglecting vocabulary, language teachers and researchers recognized the importance of vocabulary learning and started to investigate ways of promoting it (Read, 2000, p. 1).

There are two general viewpoints about vocabulary learning. One of them is represented by incidental vocabulary learning, which according to Hulstijn (2001), is "the learning of vocabulary as the by-product of any task not explicitly geared to vocabulary learning" (p. 271). The origin of this view is the Default Hypothesis, which proposes that most words are learned through exposure to language input because the number of words to be learned is too great to be accounted for by instruction (Laufer, 2010, p. 15). From this viewpoint, teachers should encourage learners to infer the meanings of words from contexts and when learners fail to do so, they should tolerate vagueness and wait for future exposure to the same words (Boers & Lindstromberg, 2008, p. 3).

The other approach is intentional vocabulary learning, which is accomplished by completing "any task aiming at committing lexical information to memory" (Hulstijn, 2001, p. 271). This view has received support from scholars who consider intentional vocabulary learning as an essential part of language learning. For instance, Coady (1997, pp. 230-232) states that beginners are not able to learn words through mere exposure to input because they do not know enough words to comprehend the input. He suggests teaching 3000 most frequent words to beginners to help them comprehend general target language input and enable them to use input as a source of learning. Referring to previous research, Richards and Renandya (2002, p. 260) mention that although incidental vocabulary learning plays a major role in advanced learners' vocabulary, intentional learning considerably contributes to vocabulary development at lower levels. Laufer and Rozovski-Roitblat (2011) also found out that intentional activities lead to longer retention compared to incidental activities and they are more effective for recognition of forms and meanings. Nation (2013, p. 95) elaborates that while intentional learning of words results in an effective boost in vocabulary knowledge, incidental learning, which follows direct teaching, can fill the gaps in the knowledge and use of the learned words. Referring to the results of previous studies, Gonzalez-Fernandez and Schmitt (2017, p. 288) state that incidental vocabulary learning, although effective, is slower than intentional learning and its uptake rate is uneven compared to intentional learning. They emphasize that for vocabulary learning, incidental learning is not enough, and intentional learning is also required.

Laufer (2010, p. 16) also states that the impact of receiving input, especially in the form of reading, on vocabulary learning should not be overestimated. Referring to empirical evidence, she elaborates that if learners are left to learn vocabulary through reading, they do not always notice unknown words in the input. Even if they notice the words, they are not always able to guess their meanings from the contexts. Moreover, guessing the meanings of words does not necessarily result in retention of the words. She adds that learners should be

exposed to words in contexts repeatedly to pick up the words, but for repetitive exposure to occur, there should be enough input, which is hardly possible in contexts where English is a foreign language (EFL contexts). As an alternative to the Default Hypothesis and incidental vocabulary learning, Laufer (2010, p. 17) proposes word-focused instruction that according to Laufer and Rozovski-Roitblat (2015), can compensate for lack of input in EFL learning contexts. Her proposition is based on form-focused instruction (FFI) in the area of grammar that, as Laufer (2010, p. 17) mentions, can be applicable to vocabulary as word-focused instruction. In the area of grammar, FFI can be either attending to linguistic elements during communicative activities or teaching discrete linguistic elements. The former is called Focus on Form (FonF) and the latter Focus on Forms (FonFs). In the area of vocabulary, as Laufer (2010, p. 17) elaborates, FonF is dealing with words during communicative tasks when the words are necessary for task completion. An example is looking up the meaning of words in a dictionary and then using them to communicate. FonFs, however, is teaching and rehearsing words in non-communicative tasks. An example is providing L1 translations of the target words followed by vocabulary exercises such as matching the words with their definitions or gap-filling activities, which are not connected with a communicative activity. Accordingly, the target words are the learning objectives.

It may be argued that only a limited number of words can be taught through word-focused instruction, given the time constraints in classroom contexts. While acknowledging this caveat, the value of word-focused instruction for learning some aspects of vocabulary and some specific groups of words is emphasized. Regarding learning aspects of vocabulary knowledge, Laufer (2005) believes that word-focused instruction can expedite elaboration of knowledge about words' paradigmatic and syntagmatic relations, constraints of use, and collocations. In Laufer's view, word-focused instruction also facilitates developing productive vocabulary knowledge. She reasons that the frequency of using any given word by learners in speech and writing is not sufficient for developing its productive aspect. Thus, learners may not be able to use them later productively. Word-focused instruction, according to Laufer, can trigger use of words and promote productive knowledge. It also promotes the speed of word retrieval form memory. Regarding learning specific groups of words, Laufer (2005) states that low frequency words need to be taught because their frequency of occurrence is insufficient for incidental learning to be successful and states that to build learners' competence, teaching of many low frequency words is necessary. Nation (2013, p. 94) also recommends teaching academic words in addition to 2000-3000 high frequency words, which are essential for language use.

Empirical evidence supports the effectiveness of word-focused instruction (e.g., De la Fuente, 2002; Hulstijn, Hollander & Greidanus, 1996; Laufer & Rozovski-Roitblat, 2011; Luppescu & Day, 1993; Paribakht & Wesche, 1997; Zimmerman, 1997). Word-focused instruction can also be justified theoretically. According to Laufer (2010, p. 24), it is related to several second language acquisition hypotheses, namely the Noticing Hypothesis (Schmidt, 1990), the Limited Capacity Hypothesis (Van Patten, 1990), the Pushed Output Hypothesis (Swain & Lapkin, 1995), and the Involvement Load Hypothesis (ILH). The last one, the ILH, which was proposed by Laufer and Hulstijn (2001) exclusively for vocabulary learning is explained in the following section. The other mentioned hypotheses are discussed in Chapter 2.

1.3 The Involvement Load Hypothesis (ILH)

Considering that word-focused instruction contributes to vocabulary development, designing effective word-focused tasks can be of great importance. For designing tasks, factors that contribute to task effectiveness should be identified and be incorporated into task design. Laufer and Hulstijn (2001) analyzed previous studies in the area of vocabulary learning and after post-hoc explanation of their results, they identified the effective factors for vocabulary learning and proposed the Involvement Load Hypothesis (ILH), which was based on three assumptions.

The first assumption is that retention of words depends on three factors in a task, namely need, search, and evaluation. Need is a motivational (non-cognitive) factor and search and evaluation are cognitive (information processing) factors. Need refers to learners' intention to understand or use a word. Search is learners' attempt to look for the meaning of an L2 word or to find an appropriate L2 form for an L1 meaning (by asking teachers or peers or looking up words in a dictionary). Evaluation involves comparing a given word with other words or comparing a specific meaning of a word with its other meanings to decide whether the word or the specific meaning of the word fits the context or not. Evaluation also entails decisions regarding the combination of words at the phrasal and clausal levels (Laufer & Hulstijn, 2001). According to Kim (2011), the cognitive factors, namely search and evaluation, entail attention to the forms and meanings of words. In Laufer's (2010, p. 25) view, these factors are form-focused activities in the area of vocabulary.

Not all the above-mentioned factors are necessarily present in a task designed for vocabulary learning. Their presence can also vary in degrees of prominence. The degrees of

each factor are elaborated by Laufer and Hulstijn (2001) and later by Nation and Webb (2011, p. 4) as follows:

- Need can have three degrees of prominence. If teachers ask learners to understand or use a word or if understanding a word is necessary for successful completion of a task, need is imposed externally and appears in moderate strength that is shown by (+). In this condition, motivation is extrinsic. However, if learners themselves decide to express a concept for which they lack vocabulary knowledge and decide to learn and use a specific word to express the concept, need is self-imposed and appears in a strong degree that is shown by (++). In this condition, motivation is intrinsic. It is also possible that learners do not decide to understand or use a word, or learning a word is not needed for task completion, or teachers do not ask learners to learn a word. In such circumstances, need is absent and shown by (-).
- Search, according to Laufer and Hulstijn (2001), only has two degrees of (-) and (+). If learners are given the meanings of new words search is shown by (-). If they are required to look for the meanings of words by asking others or looking them up in dictionaries search is shown by (+). Later, Nation and Webb (2011, p. 4) defined three degrees for search and stated that if the meanings and forms of words are provided, the search factor will be (-). If learners look for the meanings of L2 words (by asking teachers or peers or looking them up in dictionaries) or retrieve their meanings from memory, the search factor is moderate (+). This type of retrieval, which involves retrieving the meanings of words from memory, was called by Nation and Webb (2011) receptive retrieval. However, if learners look for or retrieve the appropriate forms of L2 words the search factor appears in a strong degree (++). This kind of retrieval was called by Nation and Webb (2011) productive retrieval.
- Evaluation has three degrees. If learners compare a given word with other words or a specific meaning of a word with its other meanings in order to find a word or a meaning that best fits a given context (e.g., completing gap-filling tasks), evaluation is moderate and is shown by (+). If the context is not given and learners are required to combine newly learned words with other words in order to make an original sentence/s (e.g., sentence-writing or composition-writing), evaluation will be strong and is shown by (++). In the absence of the mentioned conditions, evaluation is shown by (-).

The combination of the need, search, and evaluation factors with their degrees of prominence is called the involvement load, which is a motivational-cognitive construct. In a natural setting, different words can have different involvement loads. Nevertheless, in a teaching or research context, it is possible to adjust the involvement load for target words by adjusting the requirements of designed tasks. This involvement load is called task-induced involvement load. Task-induced involvement load can be shown as a number by summing up the degrees of the three factors. The number is called the involvement index (Laufer & Hulstijn, 2001).

Building upon the first assumption, the second assumption of the ILH was explained by Laufer and Hulstijn (2001) as follows. Under equal conditions, words that are processed with a higher involvement load will be retained better than the words processed with a lower involvement load. Accordingly, assumption three states that tasks with a higher involvement load are more effective for vocabulary learning than tasks with a lower involvement load.

The assumptions of the ILH regarding the effectiveness of vocabulary- learning tasks do not depend on task type (e.g., input or output tasks) and task mode (e.g., oral, aural, or visual tasks). They only depend on the combination of the involvement factors and their degrees of prominence (Laufer & Hulstijn, 2001).

1.4 The ILH Strengths and Weaknesses

According to Hulstijn and Laufer (2001), the ILH with its three factors has clear implications for vocabulary teaching. It assists teachers to design tasks with different involvement indices for a variety of purposes. Eckerth and Tavakoli (2012) point out that "the main educational appeal of the ILH is its potential instructional applicability: a formula for teachers to use to better manipulate and foster their students' vocabulary learning" (p. 244). The ILH can provide guidelines for teachers and materials designers to design tasks or to predict their effectiveness for vocabulary learning. If a task is found to be ineffective, the ILH guidelines make it possible to adapt the task by incorporating effective factors. Hulstijn and Laufer (2001) state that being aware of the construct of task-induced involvement not only helps teachers and material designers to design and adapt tasks, but also can help autonomous learners to make informed and strategic decisions about the tasks they select for vocabulary learning.

Hulstijn and Laufer (2001) suggest that for teaching words that are considered important for learners, such as academic words for university students, tasks with higher involvement indices are required. However, for easy words, it is sufficient to use tasks with low involvement indices, which do not take a long time and do not need too much cognitive

effort compared to tasks with high involvement indices. As Schmitt (2010, p. 53) states, learning a concrete word, which refers to an entity experienced by the senses (De Groot, 2006), is easier than learning an abstract word. Nouns may be easier to learn than verbs (Ellis & Beaton, 1993 a). In addition, the more a foreign language word conforms to phonological and orthographic patterns of learners' L1, the easier it is to learn (Ellis & Beaton, 1993 a). Hence, for teaching concrete words, nouns, frequent and short words, and words with familiar features, it may not be necessary to use tasks with higher involvement indices that take a great deal of time and effort.

Research has shown that the ILH is a good approach for analysing and predicting the effectiveness of vocabulary-learning tasks (e.g., Huang, Willson & Eslami, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2011). The ILH strength is its simplicity, which provides an easy testable way of analysing vocabulary-learning activities (Nation, 2013, p. 100; Nation & Webb, 2011, p. 5). Considering that part of vocabulary knowledge may need to be taught, a simple framework such as the ILH may provide a convenient way of designing and analysing word-focused tasks for language teachers. However, the hypothesis has some shortcomings that have been addressed in the literature.

The first defect is uncertainty about the weight of the factors for vocabulary learning, or as Laufer and Hulstijn (2001) mentioned, their importance for vocabulary learning. When Laufer and Hulstijn (2001) proposed the hypothesis with the hope of stimulating theoretical and empirical work in the area of vocabulary learning, they gave equal importance to the three factors. However, as they stated, the factors may not have the same weight for vocabulary learning. They suggest that the weight of search may be lower than evaluation and called for studies to investigate the weight of the factors for vocabulary learning. Later, Keating (2008) stated that dictionary search itself may entail some degrees of evaluation. He elaborated that when learners look up the meanings of words in a dictionary, their focus on form is greater than when they read glosses. Glosses encourage only superficial attention to new words whereas searching for the meanings entails greater focus. After that, Kim (2011) argued against giving the same degrees of -, + and ++ to the factors and treating them in the same way in terms of their impact on learning and the part they play in the overall involvement index of a task. Her presumption was that strong evaluation may induce higher involvement than the other factors. Kim (2011) also called for studies to investigate the contribution of the factors to vocabulary learning in order to refine the weights of the factors and the degrees given to them. Since then, studies such as Zou (2012), Bao (2015), and Zou (2017) have questioned the weights and degrees given to the factors. Zou (2012) suggests three degrees for search,

including zero, moderate, and strong, to take into account the processes involved in search such as inferencing and dictionary consultation, and four degrees for evaluation, including zero, moderate (for activities such as cloze exercises), and strong (for sentence-writing), and very strong (for composition-writing). Later, Zou (2017) suggested the same degrees for the evaluation factor. Bao (2015) argues against giving a moderate degree to externally imposed need and a strong degree to self-imposed need and stated that the current distinction might obscure the subtle difference between the strength of need induced by different vocabulary-focused activities.

The second shortcoming is uncertainty about the impact of distribution of the factors (Kim, 2011; Laufer & Hulstijn, 2001). In two studies carried out by Laufer (2003) and Hu and Nassaji (2016), it was found out that tasks with the same involvement indices in which the factors were distributed differently were not equally effective for vocabulary learning. Since these findings contradict the ILH assumptions, more evidence may be required regarding the effect of distribution of the factors.

Hu and Nassaji's (2016) study also revealed another defect in the ILH. As they explained, the ILH weakness is the limited range of scores it gives to tasks. The highest involvement index the ILH gives to a task is 5 (++ need, + search, and ++ evaluation). For this reason, the difference between tasks with high and low involvement indices may not be large enough to allow for predicting task effectiveness.

The above paragraphs gave an overview of the strengths and weaknesses of the ILH. The defects explained above, including uncertainty about the importance of the factors for vocabulary learning, uncertainty about the impact of the distribution of the factors, and the limited range of scores the ILH gives to tasks, provided the basis for the aims of the present research. The in-depth explanation of the ILH factors and studies in the area are presented and discussed in Chapter 2.

1.5 Research Aims

Considering the shortcomings explained in the previous section, it may be possible to improve the ILH. To this end, the present study aims to investigate the contribution of each degree of each factor to vocabulary learning. The need factor of the ILH has not received enough attention in the literature. Except for Lee and Pulido (2017), which interpreted their results by expanding on the need factor, and Bao's (2015) study, which recommended refining the

degrees given to need, studies in the area of the ILH mostly investigated the search and evaluation factors. Thus, the present study aims to focus on different degrees of need and account for their contribution to vocabulary learning.

The present study also aims to investigate the contribution of search and evaluation to vocabulary learning from a point of view that is different from those of previous studies. As mentioned before, Nation and Webb (2011, p. 4) added the concept of retrieval to search and suggested three degrees for this factor. Based on their definition, the absence of search, shown by (–), refers to a condition where no dictionary use or no retrieval occurs. Moderate degree (+) is given to search when learners look for the meanings of L2 words (by asking teachers or peers or looking them up in dictionaries) or when they retrieve the meanings of words from memory (receptive retrieval). Strong degree (++) is given to search when learners look for the appropriate forms of L2 words or when they retrieve the forms of words from memory (productive retrieval). However, the present study attempts to separate the two concepts of search (dictionary consultation) and retrieval, given that they involve different processes. As Desrochers and Begg (1987) state, memory retrieval involves three stages: trace contact, which is recognizing relevant memory content, trace use, which is making the content available, and trace decoding, which is translating memory content into response (p. 65). However, dictionary search during task completion, as Nation (2013, p. 419) explains, involves getting information from the context, finding the dictionary entry, and choosing the most suitable sense of the word that fits the context. In addition, memory retrieval is retrieving already encountered information while dictionary search, as a way of learning new words, is about finding new information. Thus, in the present study, search was solely considered as dictionary search and retrieval that is either in the form of productive memory retrieval or in the form of receptive memory retrieval was considered as a type of evaluation.

Considering retrieval as a type of evaluation is justifiable using the ILH assumptions and Nation and Webb's (2011, p. 4) proposal for giving degrees to the search factor. Based on the ILH assumptions, gap-filling tasks in which learners need to decide which word fits a given context involve evaluation. However, the degree of evaluation still depends on the type of retrieval. If learners need to fill a gap in a sentence by retrieving the form of the words from memory, as Nation and Webb (2011) state, they are involved in *productive* retrieval. On the other hand, if they need to choose from several options a form that fits a given context, which involves retrieving the meanings of words from memory, they are involved in *receptive* retrieval. In the former condition, the task induces strong evaluation and, in the latter, moderate evaluation. The given degrees of evaluation are in accordance with Nation and Webb's (2011)

view of giving a higher involvement index to productive retrieval. There is also evidence that productive retrieval is more difficult than receptive retrieval (Ellis & Beaton, 1993 b; Laufer, Elder, Hill & Congdon, 2004; Mondria & Wiersma, 2004, p. 96) and is more effective for vocabulary learning. Thus, in this study, for gap-filling activities, if learners were given the forms of the words and need to retrieve only the meanings of the words (receptive memory retrieval), the activities would induce moderate evaluation. However, if learners had to retrieve forms for meanings (productive memory retrieval), the activities would induce strong evaluation. Considering that sentence-writing and composition-writing also induce strong evaluation (based on the ILH assumptions), the new definition of retrieval, as a type of evaluation, makes it necessary to investigate the difference between productive retrieval, sentence-writing, and composition-writing, which all induce a strong degree of evaluation. If any difference is detected between them in this study, they can be given different degrees, which results in extending the evaluation factor.

Investigating the difference between sentence-writing and composition-writing can also be important from another point of view. The ILH does not account for any difference between sentence-writing and composition-writing and considers both as the activities inducing strong degrees of evaluation. However, previous studies reached contradictory conclusions regarding the similarity between sentence-writing and composition-writing. Kim (2011) concluded that sentence-writing and composition-writing are similar in terms of their impact on vocabulary learning. Laufer (2005) also claims that sentence-writing and composition-writing induce the same level of involvement and thus, have the same impact on vocabulary learning. However, studies such as Keating (2008), Zou (2012), Zou (2017) found out that these two activities do not affect vocabulary learning similarly. Keating (2008) argues that writing composition that involves producing connected discourse may require more elaborative processing that results in higher vocabulary retention compared to sentence-writing. Considering this contradictory evidence, investigating the difference between sentence-writing and composition-writing may also deepen understanding of the effectiveness of the evaluation factor for vocabulary learning.

Based on the above-mentioned information, for investigating the involvement factors in the present study, different types of evaluation will be considered: receptive retrieval (+), productive retrieval (++), sentence-writing (++), and composition-writing (++). The reason is that productive retrieval, sentence-writing, and composition-writing are all believed to induce a strong degree of evaluation. Thus, they need to be distinguished by their types not their degrees. However, when need and search are considered, they are differentiated by their degrees of prominence, which are -, +, and ++ for need and – and + for search.

Investigating the weight of each degree of need and search and each type of evaluation might remove the first shortcoming of the ILH, which is uncertainty about the contribution of the factors to vocabulary learning. This might further contribute to removing the second shortcoming of the hypothesis, which is uncertainty about the impact of the distribution of the factors on vocabulary learning. In addition, if this study could determine the effectiveness of each degree of each involvement factor, it could be possible to give more degrees to the factors and thus, the range of indices the ILH gives to tasks would be wider. This might improve the ILH potential in differentiating between tasks and might remove the third shortcoming of the ILH, which is the limited range of scores it gives to tasks. All the mentioned modifications would lead to the development of the ILH and increase its accuracy for designing effective vocabulary-learning tasks and judging their effectiveness.

1.6 Context of the Study

The data was collected in Iran where the official language is Farsi and English is a foreign language. The opportunity to interact with a native speaker is rare or inexistent. Iran is a multi-ethnic country and people of different ethnicities, depending on which part of the country they live, may speak one of the seven languages of Azari, Kurdish, Lori, Mazandarani, Gilaki, Balochi, and Arabic. In Tehran, where this research was carried out, the majority of the population are Farsi speakers, although inhabitants may still speak one of the other languages depending on their background.

English is taught as an obligatory subject in secondary schools and universities for the purposes of reading comprehension and translation, not communication. Grammar Translation and Audiolingual are the dominant methods. English is also taught in private language institutes where the priority is given to communication and Communicative Language Teaching (CLT) is the dominant method of teaching. It makes the English courses offered by language institutes fundamentally different from the courses offered by schools and universities. Iranian learners take courses in private language institutes to compensate for the shortcomings of language teaching in the public institutions. All teachers in Iranian language institutes, schools, and universities are non-native speakers.

English is the most popular foreign language for Iranian language learners. Consequently, language institutes seldom have courses for teaching languages other than English. Although English learning in Iran is not without obstacles, such as paying high tuition fees of private language institutes and lack of interaction with native speakers, Iranian English

learners are motivated to interact with the wider world and have recognized the importance of English as an international language, which can facilitate their interaction.

1.7 Thesis Structure

Chapter 1 has provided an introduction to the thesis and explained the place of vocabulary in language learning and previous and current trends in vocabulary teaching. The Involvement Load Hypothesis, which is the focus of the thesis, has been presented and its strengths and weaknesses have been discussed. Building upon the weaknesses of the hypothesis, the research aims have been elaborated. At the end of Chapter 1, the context of the study has been provided. Chapter 2 will discuss the literature including aspects of vocabulary knowledge, tasks and vocabulary learning, which will provide the basis for material design in this study, and word-focused instruction and its theoretical underpinnings, one of which is the ILH that is the focus of this study. This chapter then will go into more detail about the ILH by explaining its components and providing information about ways other than the ILH for analysing vocabulary-learning tasks. Previous research in the area including those that provided evidence regarding the effectiveness of the factors and those that checked the assumptions of the hypothesis or compared it with another framework will be discussed. At the end of the chapter, the gaps in the literature will be highlighted and research questions will be presented. Chapter 3 comprises three parts. The first part will describe the designing of the teaching and testing materials. The second part will provide detailed information about the pilot study, which resulted in modifications to the materials. The third part will elaborate on the main study and its data collection procedures. In Chapter 4, data analyses and the results of the study will be presented. Chapter 5 will discuss each research question separately and will interpret the results in the light of theoretical bases (e.g., ILH, the Retrieval Effort Hypothesis, and the hierarchy of difficulty of vocabulary knowledge) and previous research. Chapter 6 will summarize the research findings and will explain the contributions of research to theory and practice. Limitations of the study and suggestions for future research will also be presented in the last chapter.

Chapter 2: Review of the Literature

2.1 Definition of Vocabulary Knowledge

Vocabulary knowledge is a complex construct. Over the years, a variety of definitions has been provided for vocabulary knowledge. The definitions, according to Milton and Fitzpatrick (2014, pp. 1-11), take one of the three major approaches, namely the components (dimensions) approach, developmental approach, and metaphorical approach. Milton and Fitzpatrick further explain each approach by stating that the components approach characterizes vocabulary knowledge as a construct with different components such as receptive/productive knowledge and breadth/depth of knowledge. The developmental and metaphoric approaches are both based on the components approach to defining vocabulary knowledge. The developmental approach, which is simple and basic, defines vocabulary knowledge based on the components of words learned at different stages of language learning. For instance, knowledge of form precedes knowledge of collocation. The metaphorical approach introduces metaphors that comprise or describe some of the components of vocabulary knowledge. In the metaphorical approach, word knowledge cannot be understood or measured readily, but it may be better captured by metaphors. An example is defining vocabulary knowledge as a web of words (p. 10) that refers to the links between words.

In this regard, the components approach to vocabulary knowledge is the fundamental one that provides the basis for other approaches. In accordance with the components approach to vocabulary knowledge, distinctions have been made between aspects of vocabulary knowledge such as breadth/depth of knowledge, receptive/productive knowledge, and knowledge of form, meaning, and use, which are discussed in the following section.

2.1.1 Aspects of Vocabulary Knowledge

It seems that the mentioned aspects of vocabulary knowledge overlap because the concept of depth of vocabulary knowledge can be further elaborated by form, meaning, and use. Each of form, meaning, and use is also discussed along receptive and productive dimensions.

2.1.1.1 Breadth and Depth of Vocabulary Knowledge

Anderson and Freebody (1981, pp. 92-93) introduce two dimensions of vocabulary knowledge, namely breadth and depth. Vocabulary breadth or size, as Stæhr (2009) defines, is “the number

of words for which the learner has at least some knowledge of meaning" (p. 578). Vocabulary breadth may imply a learner's ability to recognize a word and retrieve its meaning or its L1 translation (Milton, 2013, p. 60). Depth of vocabulary knowledge, according to Read (1993), is "the quality of the learner's vocabulary knowledge" (p. 357) or how well a language learner knows a word. Later, Read (2004) defined three aspects of depth of vocabulary knowledge, namely precision of meaning, comprehensive word knowledge, and network knowledge (pp. 211-212). Read explained that precision of knowledge means whether a learner has limited or more elaborated knowledge of the meaning of a word. Comprehensive word knowledge entails knowledge of orthographic, phonological, morphological, syntactic, collocational, and pragmatic features in addition to semantic features. He explained network knowledge as a learner's ability to link a word to other related words and distinguish it from all the related words. Having a network of links to a word inevitably means knowing a number of words to make the links. For this reason, there are ideas (e.g., Milton, 2013, pp. 61-62; Vermeer, 2001) that vocabulary breadth and depth are closely related.

Regarding the relationship between breadth and depth of vocabulary knowledge, Meara (2009, pp. 75-77) provides explanations based on two different models, namely the list model and network model. As he mentions, in the traditional list model, words in learners' vocabulary knowledge is viewed as a list and individual words on the list may have more or less depth compared to other words. Adding new words to the list, which increases vocabulary breath, has no implication for the other words in the list. Thus, there is no link between breadth and depth of vocabulary knowledge. However, in the network model, words in learners' vocabulary knowledge are seen as nodes, which are connected together and form a network. Breadth of vocabulary knowledge is the number of nodes and depth of vocabulary knowledge is the links between the nodes. In this model, adding a new node, which increases vocabulary breadth, and adding new links, which increases vocabulary depth, have implications for the rest of the network. Based on these explanations, Meara believes that instead of breadth and depth, which does not help understanding the nature of vocabulary knowledge, terms such as size and structure or size and organization must be used. Based on the network model, the L2 lexicon is less developed and less complex than the L1 lexicon. In other words, the L2 lexicon may be smaller with simpler links between words. However, the L1 lexicon is dense and highly organized.

The relationship between breadth and depth of vocabulary knowledge is also explained by Milton (2013, p. 73). He interpreted the results of the previous studies (e.g., Milton, Wade & Hopkins, 2010; Stæhr, 2008), who indicated that depth and size dimensions were closely

related, and concluded that at least until learners gain more knowledge of vocabulary and achieve subtlety in choice and combination of words, the breadth and depth of vocabulary knowledge is the same dimension. Schmitt (2014) adds that the relationship between breadth and depth of vocabulary knowledge depends not only on learners' vocabulary size, but also on the frequency of the target words. As he explains, for higher frequency words and for learners with smaller vocabulary sizes, the difference between vocabulary size and depth is not considerable. However, for lower frequency words and for learners with larger vocabulary sizes, the difference is great.

2.1.1.2 Receptive and Productive Vocabulary Knowledge

The distinction between receptive and productive knowledge that is also called passive and active knowledge was first made by Palmer (1921, p. 2; as cited in Milton and Fitzpatrick, 2014) as the ability to recognize a word in contrast to the ability to use the word. Laufer (2005) and Nation (2013, p. 47) define receptive knowledge as the ability to comprehend input received through listening and reading, to recognize the forms of words, and to retrieve their meanings. Productive vocabulary knowledge, as Laufer (2005) and Nation (2013, p. 47) describe, is the ability to retrieve the spoken and written forms for the intended meaning. It is assumed that receptive vocabulary is larger than productive vocabulary (Melka, 1997, pp. 92-93). The relationship between receptive and productive vocabulary knowledge is seen differently by researchers. Palmberg (1987) sees active/passive distinction as a continuum in which receptive knowledge of vocabulary gradually develops into productive. Read (2000, p. 154) elaborates that after learners encounter a new word for the first time, they may not remember it later unless they see or hear it again. Gradually, after gaining more knowledge of the meaning, form, and use of the word, learners can use the word. As Read (2000, p. 154) mentions, it may not be possible to recognize a level of knowledge at which receptive knowledge of words develop into productive.

Meara (1990) argues that receptive/productive knowledge is not a continuum, but a dichotomy and receptive vocabulary is qualitatively different from productive vocabulary. In his article, he did not provide empirical evidence and to explain his view, he applied a Graph Theory, which is used to describe real phenomena by presenting certain relationships and processes as a system of nodes linked together by lines. Meara considered each word a learner knows as a node, which is connected to other words by associations (lines). Thus, the learner's vocabulary knowledge can be seen as a network. He maintained that productive vocabulary is

easily accessed from anywhere in the vocabulary network, which means it comes to learners' mind through the links (associations) each word has with other words. Productive knowledge of words makes them easily accessible and allows easy access to other words linked to them. However, receptive vocabulary is the part of the network which cannot be reached from other parts. What activates receptive knowledge of words is an external stimulus such as hearing or seeing the forms of words. As Meara states, receptive knowledge of words is not accessible for learners without such stimulus. In other words, learners can recognize vocabulary receptively when they see or hear it but cannot bring it to mind by themselves. Thus, in Meara's view, the distinction between receptive and productive vocabulary is not a gradual one and receptive vocabulary can be made productive by explicit teaching of associational links between new words and known words.

Irrespective of the nature of the distinction between receptive and productive vocabulary, Read (2000, pp. 154-155), Laufer and Goldstein (2004), and Nation (2013, p. 47) think that the terms receptive and productive are misleading. According to Read (2000), the terms receptive and productive vocabulary knowledge are too broad and when it comes to vocabulary assessment, it is not clear which specific ability each one refers to. Read (2000, pp. 155-156) introduces the terms recognition and recall to explain the distinction. For word recognition, as Read explains, the learners are presented with the target words and they are supposed to show their understanding of the meanings. For recall, the learners receive some stimulus and they are required to recall the form of the words. Laufer and Goldstein (2004) mention the problem with the distinction by saying that there are productive features in receptive knowledge. An example is translation into L1 that can be deemed as evidence of both receptive and productive knowledge. It is considered receptive knowledge because it involves demonstrating comprehension of meaning not retrieving the forms of words. It can also be productive knowledge since providing the meanings involves production.

To avoid confusion, Laufer and Goldstein (2004, pp. 405-406) propose an alternative view of vocabulary knowledge that is called strength of knowledge based on two dichotomous distinctions:

- supplying the form for a given concept in contrast to supplying the meaning for a given form
- recall in contrast to recognition (of form or meaning)

As Laufer and Goldstein (2004) elaborate, the first distinction emphasizes that the knowledge of those who can retrieve the forms of words is different from the knowledge of those who can only retrieve the meanings of words. While the former type is called active knowledge, the latter type is considered passive knowledge. The second distinction stresses that the knowledge of those who can recall the forms or meanings of words is different from the knowledge of those who can only recognize the meanings of words. The above-mentioned distinctions lead to defining four degrees of vocabulary knowledge, namely active recall, passive recall, active recognition, and passive recognition. In Laufer and Goldstein's (2004) study, active recall of words was measured by providing the L1 translation of words and asking the learners to supply the L2 target words. For measuring passive recall of words, the learners had to provide the L1 translation for each L2 word. To test active recognition, the L1 translation of each target word was given and the learners had to choose the target word from four options. For measuring passive recognition, each target word was given, and the learners had to choose its meaning from four options in the learners' L1.

The same dichotomous distinctions were discussed in Laufer, Elder, Hill, and Congdon's (2004) study. They measured active recall of words by asking learners to provide the L2 words for the L2 definitions. The first letter of each target word was given to avoid recalling words other than the target ones with the same meanings. For measuring passive recall, the learners were required to show their understanding of the meanings of the target words that appeared in incomplete sentences. The learners' task was to complete the sentences that could not be done successfully without understanding the meanings of the target words. Tests of active recognition and passive recognition were both in the multiple-choice format. In the former, the learners saw the definitions of the target words and for each definition; they needed to choose the target word from four options. The latter was the opposite. The learners had the target words, and for each target word, they were supposed to choose a definition from four options.

Laufer and Goldstein (2004) and Laufer el al. (2004) elaborate that active knowledge is superior to passive knowledge because retrieving L2 word forms that are mostly different from L1 words, with the exception of cognates, involves a complex process. Nevertheless, retrieving L1 meanings of L2 word forms that mostly refer to the concepts common to many languages is less complicated. Laufer et al. (2004) also mention that recall is a more advanced type of knowledge compared to recognition, given that recall requires a better memory trace of words. In sum, active recall is the highest degree of knowledge and the lowest one can be either passive or active recognition. Based on these explanations, Laufer and Goldstein (2004) and

Laufer et al. (2004) introduce the hierarchy of difficulty based on which active recall is the most difficult and the strongest degree of knowledge. The second position on the hierarchy of difficulty belongs to passive recall, which is followed by active recognition and passive recognition.

2.1.1.3 Knowledge of Form, Meaning, and Use

Nation (2013, pp. 48-50) defines vocabulary knowledge at the most general level involving knowledge of form, meaning, and use as shown in Table 2.1. Knowledge of form is categorized into knowledge of spoken form, written form, and words parts (affixes). Knowledge of meaning is divided into knowledge of connections between forms and meanings, knowledge of connotations, including concepts and referents, and knowledge of associations. Knowledge of use includes knowledge of grammatical functions, collocations, and constraints of use. Each of these three aspects is further defined in terms of receptive and productive knowledge.

Table 2.1

What is involved in knowing a word (Nation, 2013, p. 49)

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

So far, the nature of vocabulary knowledge has been discussed. It is worth adding that in the present study, some of the mentioned distinctions have provided the bases for designing vocabulary post-tests and interpreting the results. As will be explained in Chapter 3, vocabulary post-tests, including active recall, passive recall, and passive recognition, were designed based on Laufer el al.'s (2004) and Laufer and Goldstein's (2004) definitions. The hierarchy of difficulty they proposed was also used for interpreting the results in Chapter 5.

After defining vocabulary knowledge and explaining aspects of it, teaching and learning vocabulary are discussed in the following sections. First, vocabulary learning using tasks are explained. Building upon it, the type of tasks used in the present study is introduced. Then, vocabulary instruction is discussed, which leads to justifying word-focused instruction and relating it to the Involvement Load Hypothesis.

2.2 Task and Vocabulary Learning

Task-based language teaching (TBLT) has received a great deal of attention in the last 30 years, with the aim of promoting successful language learning (Robinson, 2011). As the name implies, TBLT is a kind of language teaching that is based on the performing of tasks by learners. There is no single definition for a task. The literature on language learning gives a variety of definitions. Skehan (1998) states that a task is an activity that meets four criteria: "Meaning is primary. There is a goal which needs to be worked toward. The activity is outcome-evaluated and there is a real-world relationship" (p. 268). Ellis (2003) identifies six criterial features of a task some of which are similar to Skehan's definition:

> "A task is a work plan. A task involves a primary focus on meaning. A task involves real world process of language use. A task can involve any of the four language skills. A task engages cognitive processes. A task has a clearly defined communicative outcome" (pp. 9-10).

Ellis further defines a task based on four key criteria (Ellis, 2009, 2012, p. 198):

- The focus of a task is on meaning where learners are involved in semantic and pragmatic processing of meaning.
- There should be a gap such as expressing ideas, conveying information, and deducing meaning.
- For successful completion of a task, learners need to rely on their own linguistic and non-linguistic resources.
- Task has a predetermined outcome that is different from language use.

These four criteria can provide the basis for differentiating between a task and other kinds of exercises. Exercises other than tasks do not meet all these criteria.

A task-based lesson can have three phases of pre-task, main task, and post-task, but the only obligatory phase is the main task and others are optional. A task can be completed by learners working individually or working in a whole class, pairs, or groups (Ellis, 2009). Ellis (2003, p. 141) elaborates that tasks can be unfocused or focused depending on whether or not they require learners to process specific linguistic features. Focused tasks have the aims of encouraging the use of the target language communicatively and of specific, predetermined linguistic features.

Ellis (2009) also mentions that tasks do not solely involve production and interaction. Tasks can either involve learners engaging in writing and speaking, which are called output-based tasks, or be input-based and provide new linguistic materials through listening and reading. The latter type has the potential to provide sufficient input, including vocabulary. Input-based tasks are those that, as Shintani (2012) states, direct learner's attention to linguistic materials in the input through listening and reading without the need to have production. Shintani (2012) adds that there are two types of input-based tasks, namely enriched input tasks and comprehension-based input tasks. In the former type, learners are exposed to input containing target words for learning. However, the learners are not required to show they have processed the input successfully. In the latter kind of input tasks, in addition to being exposed to input, the learners are required to show that they have processed the input. In this type of task, an outcome should be achieved that is conditional upon successful input comprehension (Shintani, 2012).

In this study, as will be explained in Chapter 3, the latter type of input-based tasks (comprehension-based input tasks) were used to present the target words to the participants. In Chapter 3, the reason behind using this type of task is discussed.

2.3 Vocabulary Instruction

As mentioned in Chapter 1, two general ways of vocabulary acquisition have been discussed in the literature, namely incidental and intentional learning. The distinction between the two has been defined differently by scholars. At the operational level, it is defined in terms of forewarning learners of upcoming tests. If learners are aware of a retention test that follows any activity or task, learning is called intentional; otherwise it is incidental (Eysenck, 1982, p. 198; Hulstijn, 2003, p. 356; Hulstijn, 2005). The distinction is also defined based on the

presence or absence of intention to learn words (Bruton, Lopez, & Mesa, 2011; Hulstijn, 2001). In this sense, learning words is incidental when there is no intention to commit them to memory or when they are not the main focus of activities (Ender, 2016). Under such circumstances, vocabulary is learned as a by-product of doing other activities (Hulstijn, 2001). Examples are picking up unknown words while reading or listening (Hulstijn, 2013, p. 2632). However, intentional learning occurs when deliberate attempts are made to learn new words (Hulstijn, 2005), which involves learning through direct instruction and use of vocabulary learning strategies (Gonzalez-Fernandez & Schmitt, 2017, p. 289). Examples are gap-filling activities, matching the forms of words with their definitions, choosing the correct meaning from multiple options, and flashcards (Gonzalez-Fernandez & Schmitt, 2017, p. 289).

Later, it was suggested that retention of words does not depend on the presence/absence of an upcoming test or of an intention to learn, but it depends on the level of processing (e.g., Craik & Lockhart, 1972; Craik & Tulving, 1975; Hulstijn, 2001.; Laufer & Hulstijn, 2001; Schmitt & Schmitt, 1995). Processing information at a higher level or more elaborately, which leads to higher retention, means paying careful attention to the features of words, such as meaning, semantic relations, and grammatical category. At the same time, processing at a lower level or less elaborately implies paying attention to only one or a few features of words (Hulstijn, 2001). Drawing on the level of processing approach, Hulstijn (2001) explains that both intentional and incidental learning conditions can entail information processing at a higher level, which results in better retention, or at a low level, which leads learners to process new words superficially or even to skip the words. Thus, any task that involves a high level of processing promotes vocabulary retention irrespective of whether it is counted as incidental or intentional learning condition. Thus, in Hulstijn's (2001) view, the distinction between incidental and intentional learning is not theoretically sound as long as the level of processing is involved.

In this regard, it is recommended that learner's vocabulary learning be fostered by drawing learners' attention to words in order to provoke higher levels of processing (Hulstijn, 2001; Laufer, 2001). In Laufer's (2001) view, it can be done through word-focused instruction that refers to doing something with new words, including looking up the meanings of words in a dictionary, using glosses, sentence completion, writing sentences, and compositions writing. Subsequent research has supported her view by showing that tasks that supplement reading with word-focused instruction are more effective for vocabulary learning compared to those without such instruction (e.g., Nation, 2015; Sonbul & Schmitt, 2009; Webb, 2005).

2.4 Theoretical Rationales for Word-focused Instruction

Word-focused instruction is related to several hypotheses in the area of language learning, including the Noticing Hypothesis, the Limited Capacity Hypothesis, the Pushed Output Hypothesis, and the Involvement Load Hypothesis (Laufer, 2010, p. 24).

2.4.1 The Noticing Hypothesis

As Richards and Renandya (2002, pp. 157-160) mention, researchers have distinguished several stages in the language learning process. The first stage, input, is any language resource that is used to begin the learning process. The second stage is intake, which is the part of input that is comprehended and attended to by a learner. The third stage, acquisition, is a process of incorporating new information into interlanguage. The other stages are access and output, which include access to interlanguage during communication and having production in the target language respectively. Schmidt (1990), in his Noticing Hypothesis, states that for input to be intake, it should be noticed by learners. In other words, learners progress from the first stage to the second stage of learning if they notice the features of language. In Schmidt's view, noticing is the essential and sufficient condition for learning that is transforming input to intake. However, as Schmidt (2001, p. 23) points out later, many features of input remain unnoticed due to being non-salient and infrequent. One such feature, according to Mondria and Wit-de Boer (1991), is vocabulary. The reason is that some of the words are not essential for understanding the overall meaning and, if learners' attention is not drawn to them, they may remain unnoticed.

2.4.2 The Limited Capacity Hypothesis

Van Patten (1990) postulates that "humans are limited capacity processors and the amount of conscious attention available for the processing of incoming data is finite at any given moment" (p. 295). He maintains that when learners process language input consciously for meaning, they are unable to process linguistic forms consciously at the same time because they have a limited capacity for these simultaneous processes. As a result, learners attend to meanings and ignore forms in the input. According to Laufer (1998), this explains why learners do not notice words in the input. She highlights the importance of drawing learners' attention to form (vocabulary). According to Laufer, even for productive skills such as writing, teachers should focus on form because learners tend to use simple and frequent words to express meaning.

2.4.3 The Pushed Output Hypothesis

Swain and Lapkin (1995) argue that output has an essential role in learning a language because it triggers noticing. They explain that producing the target language helps learners to recognize their linguistic problems. Noticing the problems may in turn push learners to modify their output. Since then, studies have supported the role of output in vocabulary learning. For instance, Pica, Holliday, Lewis, and Morgenthaler (1989) indicated that learners internalized new linguistic forms through modified output. De la Fuente (2002) concluded that output plays an essential role in learning productive aspects of vocabulary knowledge, while Hanaoka (2007) found that during output, learners notice their lexical problems and feel the need to search for solutions. Laufer (2010, p. 25), emphasizing that output has a role to play in vocabulary learning, states that teachers should facilitate vocabulary learning by drawing learners' attention to words and asking them to modify their output by selecting different words or choosing more suitable words for a context.

2.4.4 The Involvement Load Hypothesis

The nature of cognitive processes involved in learning is explained by Craik and Lockhart's (1972) Depth of Processing Theory, which states that learning depends on the quality of mental activity at the time of learning in a way that deeper processing entails better learning. However, Depth of Processing Theory could not account for what constituted a level of processing and how it was possible to determine the depth of a level. Craik and Lockhart's (1972) theory was unable to answer these questions because the concept of the depth of processing lacked an operational definition (Hulstijn, 2001; Hulstijn & Laufer, 2001; Kim, 2011). Later, the Involvement Load Hypothesis (ILH) proposed by Laufer and Hulstijn (2001) provided a measurable definition for depth of processing and made it possible to apply this notion to L2 vocabulary learning (Kim, 2011). The ILH proposed a motivational factor (need), and two cognitive factors (search and evaluation) as the effective factors in any vocabulary-learning task and gave them degrees of prominence, thereby allowed quantification of the cognitive processes involved in learning new words. The cognitive and motivational factors are explained in the following sections.

2.4.4.1 The Cognitive Factors

The ILH introduces two cognitive factors, namely search and evaluation, as effective factors for vocabulary learning. In the present study, as mentioned in 1.5, search only refers to

dictionary look up activity. Evaluation can be induced by either retrieval or writing. In the following sections, the cognitive factors of the ILH are explained.

2.4.4.1.1 Dictionary Search

Dictionaries are helpful for comprehension, production, and learning (Nation, 2013, p. 414). According to Nation, dictionaries help comprehension when learners look up words during reading and listening and when they confirm guesses from context. Dictionaries help production when learners look up words during speaking and writing, when they look up words' features (spelling, pronunciation, inflections, collocations, etc.), and when they find synonyms or try to correct an error. Dictionaries help learning when learners choose unknown words to learn and when they enrich their knowledge of words.

Previous studies also indicated that dictionary use has a positive impact on vocabulary learning (e.g., Knight, 1994; Luppescu & Day, 1993). Considering the results of previous studies in the area of vocabulary learning, Laufer and Hulstijn (2001) mentioned dictionary search as one of the effective factors for vocabulary learning and introduced it as one of the components of the ILH. In their view, dictionary search induces some degree of involvement and promotes vocabulary learning.

2.4.4.1.2 Evaluation

2.4.4.1.2.1 Retrieval

Retrieval is "accessing stored information" (Roediger & Guynn, 1996, p. 197). When it comes to vocabulary learning, learners first notice new words in the input and comprehend their meanings (Nation, 2013, p. 107). Then, they need to retrieve them from memory in order to use and understand them later, which strengthens the memory for the words (Barcroft, 2007; Nation, 2013, p. 107). As Nation (2013, p. 107) mentions, retrieval can occur in receptive and productive directions. It is receptive if learners perceive the form and retrieve its meaning during reading or listening. It is productive if learners retrieve spoken or written forms to communicate the meanings of words. Nation adds that retrieval does not occur if both meaning and form are provided together. In the present study, gap-filling activities in which the learners had to either retrieve the meanings of words or the forms of words to fill in the gaps of the sentences are examples of receptive and productive memory retrieval respectively.

Regarding the effectiveness of each type of retrieval, the Retrieval Effort Hypothesis (Pyc & Rawson, 2009) states that the effectiveness of successful retrievals depends on the level of difficulty of retrievals in a way that difficult retrievals enhance memory more than easy retrievals. Based on the ILH assumptions, retrieving the form to express meaning (productive retrieval) is more difficult than retrieving the meaning (receptive retrieval). Evidence has also shown that productive retrieval, which involves greater effort, is more effective for vocabulary learning (e.g., Ellis & Beaton, 1993 b).

2.4.4.1.2.2 Writing (Generative Use)

Another process that leads to remembering words is generative use, which is called creative use by Nation (2013, p. 110). Generating target items develops the cognitive processes involved in retrieving words, which may facilitate vocabulary learning (Barcroft, 2007). Generative use can be receptive or productive. A productive form of generative use occurs when learners use previously met words in a way that is different from their previous meeting with the words. Activities such as sentence-writing and composition-writing used in the present study are examples of the productive form of generative use. In contrast, a receptive form of generative use occurs when learners meet previously met words in new contexts that are different from their previous encounter with the words. An example of receptive generative use is meeting already met words in listening and reading (Nation, 2013, p. 110). As Nation and Webb (2011, p. 9) state, productive generative use is more difficult than receptive generative use.

2.4.4.2 The Motivational Factor

In addition to the cognitive factors, which allow quantification of the cognitive processes involved in vocabulary learning, the ILH has a motivational factor, need, which allows for the quantification of motivation for vocabulary learning. Nation (2001, p. 63) emphasizes the role of motivation by highlighting that motivation and interest promote noticing, which is a prerequisite for language learning. In defining the motivational factor of the ILH, Laufer and Hulstijn (2001) refer to a statement by Gray (1999) that a human is not just an information-processing machine and that motivation and emotion affect the way humans process information. Laufer and Hulstijn apply this notion to vocabulary learning and state that learning a large number of words requires both cognitive processes and motivation. They included one aspect of motivation, need, in the ILH and pointed out that need creates tension and some

degree of tension can promote information processing and learning. The need factor Laufer and Hulstijn define is interpreted in a positive sense that is the need to achieve and the need to comply with the requirements of a task.

2.5 Ways Other Than the ILH for Analysing Vocabulary-learning Tasks

In addition to the ILH, two other ways have also been developed for designing vocabulary-focused task and analysing their effectiveness. The shortcomings of each were also recognized by previous studies, which are discussed in the following sections.

2.5.1 Nation's (2001) System

As a way of designing, analysing, and adapting vocabulary-learning tasks, Nation (2001, pp. 60-62) suggests looking at four features of tasks, including learning goals, psychological conditions, signs of learning, and design features. He elaborates, based on research evidence, that these features provide necessary conditions for vocabulary learning. The first feature, goals, can be learning all or some of the aspects of knowing a word, namely form, meaning, and use. In Nation's view, a vocabulary-learning task should provide sufficient information required for achieving a certain goal such as information about the meanings of words that can be taken from input, teacher, dictionary, or peers.

The second feature, psychological conditions, encompasses three processes that may help in remembering a word, namely noticing, retrieval, and creative use (Nation, 2001, p. 63). Noticing involves paying attention to target items that depends on words salience in input, previous exposure to words, and learners' awareness that target words may fill the gap in their language knowledge. Noticing happens when a word is explained, looked up in a dictionary, guessed from a context, or deliberately studied. Noticing entails decontextualization, which implies giving attention to a target item as an element of a language not as an element of a message. For decontextualization, a word is detached from its context to be focused on either briefly or for a longer time. A vocabulary-learning task should provide conditions for learners to attend to target words not merely as a part of a message, but as a part of the language system. Decontextualization may occur when the words are defined or negotiated. Therefore, to facilitate decontextualization, teachers can explain words or highlight them in written input and help learners to decontextualize the words by themselves. Learners may also notice a new word while they read or listen and focus on it (Nation, 2001, pp. 63-64). Although learners may notice a word through negotiation, teachers' definitions, dictionary use, or listening and reading

activities, they also need to retrieve the word to strengthen their memory of that word. Therefore, a vocabulary-learning task should provide the opportunities for retrieval that as mentioned before, can be receptive or productive. The former is retrieving the meaning by seeing or hearing a word form and the latter involves retrieving the spoken or written form of a word to communicate meaning. Creative (generative) use, as mentioned before, happens when target words are met or used after the first exposure, but in a way that is different from the previous meeting (Nation, 2001, p. 68).

The third feature, which is signs of learning, helps teachers to decide if a vocabulary-learning task needs further adaptation or not, although the presence of the signs does not necessarily show learning. Some of the conditions cannot be observed directly; however, conditions such as negotiation and generative use are observable. If a teacher sees that learners use targets words in new ways, it should be taken as a positive sign (Nation, 2001, p. 72).

The last feature is the design features of vocabulary-learning tasks that provide the conditions for learning. Design features such as the way information is presented or used (e.g., shared or split information) and types of outcome can foster the occurrence of noticing, retrieval, and creative use that are essential for vocabulary learning. Teachers should know what design features help achieving a certain learning goal. For instance, shared or split information can foster negotiation. Highlighting target words, glossing, and including them in important parts of input can encourage noticing. Asking learners to retell the story they have already read can lead to retrieval. Having knowledge of design features that facilitate achieving certain goals, teachers can create and adapt vocabulary-learning tasks (Nation, 2001, pp. 72-74). Nation (2001) summarized the above-mentioned information in Table 2.2.

Table 2.2

The Condition of learning, signs, and features in activities with a vocabulary-learning goal (Nation, 2001, p. 75)

Psychological conditions encouraging learning	Signs that the conditions are likely to be occurring	Design features of the activities that promote the conditions
Noticing a word	The learner consults a glossary. The learner pauses over the word. The learner negotiates the word.	Definition, glosses, highlighting unknown words in salient positions
Retrieving a word	The learner pauses to recall a meaning. The learner does not need to consult a dictionary or gloss. The learner produces a previously unknown word	Retelling spoken or written input
Using a word generatively	The learner produces a word in a new sentence context. The learners produce associations, causal links, etc.	Role play based on written input Retelling without the input text Brainstorming

Nation's system for observing and improving vocabulary-learning tasks has a shortcoming that is explained in the following section.

2.5.2 Technique Feature Analysis (TFA)

Nation and Webb (2011, p. 7) discuss the limitation of Nation's system by pointing out that this system does not allow for the quantification that the ILH allows. Although the ILH provides quantification, these authors discuss the limitations of this hypothesis as well and state that the hypothesis does not take into account many features that have been shown by research to be influential for vocabulary learning. For this reason, it cannot be a good framework for designing and adapting tasks. In their view, a more comprehensive set of criteria is required (p. 7). Technique Feature Analysis (TFA) proposed by Nation and Webb (2011) is in fact one of the first attempts to compensate for the shortcomings of the ILH by taking into account more factors. TFA includes 18 criteria related to motivation, noticing, retrieval, generation, and retention shown in Table 2.3. Nation and Webb's opinion is that more features can still be added to TFA.

In analysing or designing a vocabulary-learning task, each of the criteria can be 0 or 1 depending on the answer given to each question. If the answer is no, 0 will be given and if yes, the task receives 1. The highest number of points given to a task can be 18 and the same as the

ILH assumptions, the higher the index of a task, the more effective it is for vocabulary retention.

Table 2.3

The checklist for the Technique Feature Analysis (Nation & Webb, 2011, p. 7)

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

As shown in Table 2.3, motivation has three criteria. The first one stresses the importance of having a clear vocabulary-learning goal for motivating learners. The second one draws teachers' attention to task characteristics that can motivate learners. This criterion asks teachers to include such characteristics in vocabulary focused tasks. According to Nation and Webb,

challenging activities, such as those done for pleasure, and those that raise learners' awareness of successful learning, can motivate learning. The last criterion of motivation is the counterpart of the need factor of the ILH. Based on this criterion, attention to words that is the result of learners' interest is more effective than attention to words that is guided by teachers (Nation & Webb, 2011, p. 8).

The same as motivation, noticing has three criteria. The first criterion shows providing instruction that makes learners pay attention to words, by means of decontextualizing, highlighting, and glossing, may facilitate learning. The second criterion brings the importance of noticing into focus by saying that learners should notice that there is something to learn. Examples of activities that raise awareness of new vocabulary learning include using words in contexts and choosing correct words from several options. The last criterion of noticing sees negotiation as important for vocabulary learning (Nation & Webb, 2011, p. 8).

Retrieval has five criteria. The first criterion is the counterpart of the search factor of the ILH. Tasks that provide both form and meaning receive 0 because they do not encourage retrieval. The other criteria in this category provide details of different types of retrieval. Productive retrieval (providing word forms) is more difficult than receptive retrieval (providing the meanings of words). Therefore, productive retrieval receives 1 and receptive retrieval receives 0. Likewise, recall (retrieving the forms or meanings of words) is more difficult than recognition (choosing the forms or meanings of words from several options). The number of retrievals and the space between retrievals are also considered important. A higher number of retrievals makes a task more effective for vocabulary learning. If words are retrieved more than once, spacing can be more effective than successive retrieval (Nation & Webb, 2011, pp. 8-9).

Generation with three criteria emphasizes that generative use strengthens memory for target words. However, productive generative use receives 1 point as it is more demanding than receptive generative use. The last criterion of generation is applicable to tasks that involve only productive generation. Writing an original sentence induces a high degree of generation and receives 1 point. However, rewording receives no point because the original sentence puts limitations on the degree of generation (Nation & Webb, 2011, p. 9).

Retention with four criteria takes into account the way target words are presented in a task. The first criterion stresses the importance of linking the forms and meanings of words. Accordingly, only activities that provide the chance of successful linking of forms and meanings receive points. The links between forms and meanings are strengthened by retrieval. Therefore, tasks that provide both forms and meanings do not meet this criterion. Instantiation is the second criterion for giving points to tasks. Instantiation, which is providing instances of

words, is necessary for remembering the words. Instantiation can be in the form of using a word in a meaningful situation accompanied by visual presentation of the object, action, or quality it refers to. Imaging, the third criterion, is seeing or imagining a visual image related to the meaning of a word such as visualizing an example sentence containing the target word in a dictionary. The last criterion rules out the tasks that cause interference by presenting target words that are members of a lexical set such as synonyms and antonyms at the same time (Nation & Webb, 2011, p. 10).

The same as the ILH and Nation's system, TFA has some shortcomings. Nation and Webb (2011) mention that although TFA takes into account more factors that affect vocabulary learning, there are still many other factors involved in learning words that were not included. In addition, the aspects do not have equal impact on vocabulary learning and TFA does not allow for weighting the aspects. Nation and Webb also state that another weakness of TFA is the inclusion of 'new learning' under noticing. In their view, it is not easy to decide what is considered new learning. If a word is introduced, but follow-up activities focus on some unmet aspects of the word, it is not clear whether it is counted as new learning or not (p. 15).

2.6 Research in the Area of the ILH

Research in the area of the ILH has been divided into three parts. The first part summarizes studies which focused on the involvement factors and provided evidence about their effectiveness for vocabulary learning. The second part presents studies that adopted a more holistic approach and provided evidence for and against the assumptions of the ILH. The last part gives a summary of the studies that compared the ILH with the TFA and could provide valuable insights about the strengths and shortcomings of both frameworks.

2.6.1. Evidence about the Effectiveness of the Involvement Factors

Laufer and Hulstijn (2001) analysed previous studies in the area of vocabulary teaching and learning and, after detecting general patterns in data regarding task features that affect vocabulary learning, they proposed the ILH. As they mentioned, need (with three degrees), search (with two degrees), and evaluation (with three degrees) are effective factors in vocabulary-focused tasks and the higher the degree of these factors in a task, the more effective the task is for vocabulary learning. They gave an example that a task inducing moderate need (+), search (+), and no evaluation (-) has the same involvement load as a task inducing moderate

need (+), no search (-), and moderate evaluation (+). However, As Laufer and Hulstijn (2001) state, all these factors may not be equally effective for vocabulary learning and the weight of search may be lower than the weight of need and evaluation. They emphasized that the hypothesis is the first step in stimulating theoretical and empirical work and called for research to define the ILH factors precisely, to add new factors, and to investigate theoretical links between the hypothesis and theories of information processing. Since then, studies have been carried out to investigate different aspects of the ILH.

The first study was conducted by Hulstijn and Laufer (2001) themselves. It comprised two experiments with participants at different levels of proficiency to investigate whether vocabulary learning depends on task-induced involvement. The target words were practiced through three tasks with different involvement loads, including reading comprehension, comprehension plus gap-fill, and composition-writing. Learning was measured by a passive recall test in which the participants had to provide L1 translations or English explanations of the target words. The results of the experiment with learners at a higher level of proficiency fully supported the ILH predictions in that composition-writing with the highest involvement index led to superior vocabulary retention. However, the findings of the experiment with low proficiency learners only lent partial support to the ILH, showing that reading comprehension and comprehension plus gap-fill, with medium and low involvement indices respectively, were equally effective.

The results of Hulstijn and Laufer's study were questioned by Keating (2008), who stated that in their experiments the effect of time on task had not been taken into account. In Keating's view, the time it took to complete a task could affect the task effectiveness and led to misinterpretation of the results. Keating also raised other issues regarding the impact of task involvement load. First, in his view, it was not clear whether learners at the beginning level of proficiency could benefit from tasks with higher involvement load or not. Second, the effects of task involvement on productive (active) vocabulary learning were not clear, given that in Hulstijn and Laufer's study only receptive (passive) vocabulary was taken into account. In Keating's (2008) study, the ILH assumptions were investigated by comparing three tasks with different involvement loads in terms of their impacts on active and passive vocabulary learning. The tasks, including reading comprehension (no involvement), reading comprehension plus gap-fill (moderate involvement), and sentence-writing (strong involvement), were completed by the participants. Time on task was controlled by converting the participants' scores on active and passive recall tests to take into account the words learned per minute. Learning was measured by administering immediate and delayed active and passive recall tests. Keating's

study indicated that the ILH assumptions are generalizable to beginning learners and are true for both active and passive vocabulary learning. Nonetheless, the benefit of tasks with higher involvement load did not hold when time on task was taken into account. Keating's study was revealing from another point of view. Comparing the tasks used in Hulstijn and Laufer's study and in his study and the achieved results, Keating concluded that composition-writing might be different from sentence-writing in terms of its impact on vocabulary retention although the ILH had given the same weight to them.

This issue was investigated by Kim (2011), who partially replicated Hulstijn and Laufer's study by controlling for time on task. Kim's study comprised two experiments with participants at two different levels of proficiency. In the first experiment, she investigated the effect of three tasks with different involvement loads on initial learning and long-term retention of words. Three tasks included reading comprehension with marginal glosses, reading comprehension with marginal glosses plus gap-fill, and writing a composition using the target words. The Vocabulary Knowledge Scale designed by Paribakht and Wesche (1993) was used to measure learning. The results related to long-term retention of words fully supported the ILH assumptions. However, when it came to initial learning, evidence only partially supported the ILH assumptions in that the gap-fill task (with higher involvement load) did not result in superior learning compared to the reading task (with lower involvement load). Kim concluded that at least for initial vocabulary learning, the extent to which each involvement factor contributed to the overall involvement index might not be the same. Her interpretation was that, although Laufer and Hulstijn had given equal importance to the factors when determining the involvement index of tasks, the factors may not induce the same involvement in processing a word for initial learning. Kim called for studies to examine the contribution of each involvement factor to vocabulary learning. She added that giving the same values of minus, moderate, and strong to the three factors should also be justified, given that strong evaluation may be more important than the other factors. In the second experiment, Kim compared sentence-writing and composition-writing in terms of their impacts on initial learning and long-term retention of L2 words. Learning was again measured by the Vocabulary Knowledge Scale. It was shown that sentence-writing and composition-writing with the same involvement index had similar effects on initial learning and long-term retention of words, which contradicted the results obtained in Keating's (2008) study. In addition, the study indicated that the effect of task involvement on vocabulary learning is not affected by learners' levels of proficiency.

Similar to Kim, Zou (2012) questioned the degrees given to the search and evaluation factors by stating that search was given two degrees of minus and one while different ways of

finding the meanings of words (e.g., dictionary consultation and inferencing) were not taken into account. She added that the degrees of prominence given to evaluation were also not differentiated clearly. In an attempt to modify the ILH, Zou (2012) carried out a study with five groups of advanced learners. The tasks included composition-writing and sentence-writing with an equal involvement load of 3 and reading plus dictionary consultation, reading plus inferring, and reading plus cloze exercises with an equal involvement load of 2. The Vocabulary Knowledge Scale was used to measure the participants' learning of the target words. The results partially supported the ILH. On the one hand, writing tasks with higher involvement load were more effective than reading tasks with lower involvement load, which conforms to the ILH assumption. On the other hand, tasks with the same involvement load were not similarly effective for vocabulary learning. Reading plus dictionary search was more effective than the other reading tasks. Likewise, composition-writing was more effective than sentence-writing. Based on the results, Zou proposed that search be given three degrees of minus, moderate (inferencing), and strong (dictionary consultation) and evaluation be given four degrees of minus, moderate (cloze exercise), and strong (sentence-writing), and very strong (composition-writing).

Later, Zou (2017) carried out a study with a specific focus on the evaluation factor. Three tasks, including a cloze-exercise, sentence-writing, and composition-writing, were compared. Learners' receptive knowledge of the meanings of words and productive knowledge of meanings and use were measured using Folse's (2006) modified version of Paribakht and Wesche's (1997) Vocabulary Knowledge Scale. The result partially supported the hypothesis, showing that the writing tasks, which induced strong evaluation, were more effective than the cloze-exercise, which induced moderate evaluation. However, the writing tasks with the same involvement load did not have the same impact on vocabulary learning. Composition-writing was significantly more effective than sentence-writing. Based on the result, Zou emphasized her proposal to give four degrees to the evaluation factor, including no evaluation (-), moderate evaluation (+) for cloze exercises, strong evaluation (++) for sentence- writing, and very strong evaluation (+++) for composition-writing.

The studies mentioned above mostly focused on the cognitive factors of the ILH. Lee and Pulido's (2017) study might be one of very few studies that tried to expand on the need factor, although their definition of need was different from the definition provided by the ILH. In this study, need was defined as topic interest and its impact on vocabulary learning was investigated. Words were learned through reading and answering comprehension questions on two texts, one of which was rated by learners as low-interest and the other high-interest. The

results revealed that topic interest significantly affects vocabulary learning. Lee and Pulido's interpretation was that, when intrinsic interest in a text's topic is high, the need factor (the motivational factor of the hypothesis) becomes stronger and increases the involvement load of tasks. They supported the inclusion of a motivation factor in the ILH and confirmed that both motivational and cognitive factors play essential roles in vocabulary learning.

The mentioned studies mostly focused on the involvement factors and how and to what degree they are induced in vocabulary learning activities. However, there are studies that aimed to check the assumptions of the hypothesis or use the assumptions to interpret their results. These studies are discussed in the following section.

2.6.2 Evidence for and against the ILH Assumptions

Some studies directly investigated the ILH and generally supported its assumptions; each provided more insight into the hypothesis. There are also other studies that did not aim to examine the ILH, but their results were interpretable using the hypothesis assumptions. Rott's (2007) study is an example of those that indirectly examined the ILH assumptions. She investigated the effect of frequency of input-enhancement on vocabulary learning measured by active recall, passive recall, and passive recognition tests. She found out that glossing target words four times in the text (4G) or glossing once followed by retrieval in L1 and bolding twice (GR) resulted in more productive word gain compared to glossing once and bolding three times (GB). Rott interpreted the results based on the ILH assumptions and factors. She gave the same involvement index of 4 to 4G and GR and the involvement index of 1 to GB and claimed that 4G and GR with a higher involvement load led to superior vocabulary gain. With respect to assigning the involvement index to 4G, she stated that glossing the words provoked extrinsic or moderate need (+) for knowing the meanings of target words, so 4G (glossing four times) would induce the involvement index of 4 (1×4). For assigning the involvement index to GR, she mentioned that retrieval induced moderate need (+) for the target words in addition to search (+) for the meaning from memory and moderate evaluation (+) of the correct meaning, which resulted in the total involvement index of 3. Thus, in GR, where glossing with the index of 1 was followed by retrieval with the index of 3, the resulting involvement index was 4 (1+3). However, Rott's interpretation was that bolding the target words in GB was a less obtrusive intervention compared to glossing and retrieval. It directed learners' attention to the forms of words and did not necessarily trigger processing of their meanings. Rott's assumption was that the involvement index of GB was 1 as it induced need only once during glossing.

Another example is Pichette, De Serres, and Lafontaine's (2011) study, whose results could be accounted for by the ILH assumption. They compared the effectiveness of reading and writing for learning concrete and abstract vocabulary. In accordance with the ILH assumptions, the immediate cued recall test indicated that the writing task (with a higher involvement load) was superior to the reading task (with a lower involvement load), with more concrete words being recalled compared to abstract word. However, in the delayed recall test, the superiority of writing over reading faded for abstract words. As Pichette, De Serres, and Lafontaine stated, assuming that delayed recall would be a better indicator of learning compared to immediate recall, the ILH assumption that writing could result in higher vocabulary learning due to inducing a higher involvement load was supported only for concrete words not for abstract words.

In addition to the mentioned studies that indirectly examined the ILH assumptions, there are example of the studies that directly investigated the hypothesis. One of them is Nassaji and Hu's (2012) study in which three reading tasks with different involvement loads were completed by learners. In the task with a low involvement load, learners had to read a text and infer the meanings of words by choosing a word from multiple choice glosses. In the task with a moderate involvement load, learners had to infer the meanings with no options provided. Finally, in the task with a high involvement load, learners not only had to infer meanings, but also had to make derivational changes to the target words. Vocabulary Knowledge Scale was used to measure the participants' learning of the target words. The result supported the hypothesis assumptions, showing that tasks with a higher involvement load resulted in superior vocabulary learning.

Huang, Willson, and Eslami (2012) conducted a meta-analysis, including twelve studies, in the area of foreign language vocabulary learning. They found out that output tasks such as gap-filling, sentence-writing, and composition-writing promote vocabulary learning more than tasks that do not involve output. They posited that their findings support the ILH assumptions in that the superiority of output tasks is due to their higher involvement load. They also concluded that output tasks with higher involvement loads were more effective for vocabulary learning in comparison with output tasks with lower involvement loads. This meta-analysis also revealed that time on task is a factor that affects vocabulary retention in a way that the more time learners spent on a task, the more vocabulary they learned.

Another study that directly examined the hypothesis assumptions was carried out by Eckerth and Tavakoli (2012). In this study, three tasks, including reading with marginal glosses, filling gaps using a word list, and reading with marginal glosses plus writing a

composition, were completed by learners. In all the tasks, half of the target words occurred once and half of them five times to check the effect of frequency of exposure. Active and passive word recognition and recall were measured and the results indicated that word gain depended on both frequency of exposure and elaboration of word processing. With respect to word processing, Eckerth and Tavakoli concluded that the third task was more effective than the other ones and the second one was more effective than the first because of having higher involvement indices. In this study, the effect of frequency of exposure faded over time, but the effect of task involvement load remained.

Snoder and Reynolds (2019) compared Lexical Input Processing (Barcroft, 2015) and the ILH in terms of their power for predicting the learning of collocations while performing a dictogloss. Two different versions of the dictogloss were compared. In the first one, which involved semantic elaboration, learners received a glossed list of words and were asked to write an original sentence for each and share them orally with their peers. It was followed by a text that was read aloud by a teacher. The learners had to take notes of the key words and phrases, and then interact with peers to write the text based on their notes and what they remembered from the text read by their teacher. The ILH gave the index of 3 to this dictogloss. In the second dictogloss, which involved structural elaboration, the learners received a glossed list of words and were asked to write original phrases that rhymed with them and share them orally with peers. It was followed by the same read aloud and writing procedures. The ILH gave the lower index of 1 to this dictogloss. Thus, based on the ILH assumptions, the former task had to be more effective than the latter due to having a higher involvement index. However, based on Barcroft's Lexical Input Processing theory, the latter dictogloss that involved structural elaboration had to be more effective than the former one, which involved semantic elaboration. Learning of collocations was measured using productive and receptive tests. The results were contrary to the predictions of the Lexical Input Processing theory, showing that semantic elaboration was more effective than structural elaboration. However, the results conformed to the ILH assumptions in that tasks with a higher involvement load resulted in superior learning.

The mentioned studies lent support to the ILH assumptions; however, several studies cast doubt on the assumptions. A study that found contradictory evidence was conducted by Folse (2006), who investigated the effect of different types of written exercises on word retention. The tasks he used included one gap-filling exercise, three gap-filling exercises, and sentence-writing. Vocabulary knowledge scale was used to measure word retention. The result suggested that the task with the three gap-filling exercise was the most effective one. Folse claimed that multiple retrieval of the target words in tasks that induced a lower involvement

load might be more effective than deep processing of words through writing tasks that induced a higher involvement load. In his view, the number of retrievals was more important than depth of processing or the involvement load. However, the conclusion drawn from this study can be problematic because Folse did not give any involvement index to the number of times each factor of the involvement load was induced. The same as what Rott (2007) did in her study, involvement indices should be given to tasks based not only on the presence or absence of the ILH factors, but also on the number of times each factor appears. Thus, one gap-filling exercise can be given the involvement index of 3 (+ need, + search, and + evaluation). Three gap-filling exercises have the same involvement for need (+) and search (+), but because evaluation was experienced three times (1×3), this task could be given the index of 5 (2+3). The index of sentence-writing (+need, +search, and ++evaluation) was 4. Therefore, it is not surprising that in Folse's study, three gap-filling exercises were more effective than sentence-writing as it induced a higher involvement load.

Martinez-Fernandez (2008) conducted a study to examine the effect of task involvement load on learning concrete and abstract vocabulary, taking awareness into account. The reason for including awareness in the study, as she stated, was that different levels of language processing and awareness had been measured in cognitive psychology and other areas of SLA, but not in the area of incidental vocabulary learning. In her study, three tasks were designed based on the involvement factors. In the first task, a single gloss task (+ need, - search, - evaluation), learners had to read a text with translation single glosses of the target words. In the second task, a fill-in task (+ need, - search, + evaluation), the target words were deleted from the text and learners had to fill the gaps by choosing words from a list of L2 words accompanied by their translations. In the third task, a multiple-choice gloss task (+ need, + search, + evaluation), learners read the text for which translation multiple-choice glosses were provided. Awareness, which was operationalized as noticing of meaning only, noticing of form only, and noticing both meaning and form, was measured measured based on think-aloud protocols and written recall protocols. The results indicated that the fill-in task was more effective than the multiple-choice gloss task in the meaning production and recognition tests, which contradicted the ILH assumptions because the multiple-choice gloss task had a higher involvement load and should have been more effective. In addition, the multiple-choice gloss task was not different from the single gloss task, which also went against the hypothesis assumptions. For all groups, learning concrete words was superior to learning abstract words. With respect to the level of awareness, the study indicated that tasks with different involvement loads induced different levels of awareness. The fill-in task induced higher awareness

compared to other tasks, whereas the multiple-choice gloss task induced higher awareness compared to the single gloss task. Martinez-Fernandez stated that her study did not support the ILH assumptions and explained the result based on the level of awareness each task induced. She concluded that fill-in and multiple-choice gloss tasks might involve different processing, which is unrelated to the degree of evaluation. Furthermore, the evaluation factor might be more effective than the search factor in inducing higher awareness.

Another study, which questioned the assumptions of the ILH, was carried out by Bao (2015). The study investigated the effect of task type on receptive and productive vocabulary learning. Each of the five groups of learners were exposed to target words during a reading exercise and then completed one of the tasks designed based on the ILH factors, including a control task (- need, - search, - evaluation), definition (+ need, - search, + evaluation), combining (+ need, - search, + evaluation), translation (+ need, - search, + evaluation), and writing (+ need, - search, ++ evaluation). The control group had to match the meanings with some nontarget words. For the definition task, target words had to be written in front of the given definitions. For the combining task, sentences that were segmented into word combinations had to be put together. The translation task had to be done from L2 to L1 and finally, for the writing task, the participants had to write separate sentences for the given target words. The results indicated that according to the receptive vocabulary test, the definition task was superior to each of the other output tasks. In the productive test, the definition, translation, and writing tasks showed no significant difference. Bao concluded that task effectiveness does not always match the ILH assumptions because the definition, combining, and translation tasks had the same involvement index and had to be similar in terms of their effectiveness for vocabulary learning. In addition, the writing task had the highest involvement index and had to be the most effective one. Bao mentioned that the same involvement load does not necessarily result in similar word gain and higher involvement load does not necessarily cause higher word gain. He argued that factors other than the involvement load such as contextual clues and word frequency might affect task effectiveness. The definition task probably provided learners with strong contextual clues and the high frequency at which the target words were presented in this task might induce strong need. Based on this conclusion, Bao argued against giving a moderate degree to externally imposed need and a strong degree to self-imposed need. In his view, the current distinction might obscure the subtle difference between the strengths of need induced by different vocabulary-focused tasks.

The above-mentioned studies either supported the ILH or found evidence against the hypothesis assumptions. There are also some studies that compared the ILH with TFA, which provided more insight into the ILH. These studies are discussed in the following section.

2.6.3 Comparing the ILH with TFA

A study done by Hu and Nassaji (2016) compared the ILH with three factors and the Technique Feature Analysis (TFA) with five components in terms of their ability to predict task effectiveness for vocabulary learning. Hu and Nassaji chose four tasks and ranked them based on the ILH and TFA rating systems. Then, they investigated which approach could provide more accurate predictions. The tasks included reading plus multiple-choice items, reading plus choosing definitions, reading plus fill-in-the gaps, and reading plus rewording the sentences. The first, second, and fourth tasks were given the indices of 3 and 6 by the ILH and TFA respectively. However, the third task was given the index of 2 by the ILH and 7 by the TFA. A receptive vocabulary test was administered to measure the participants' learning. Some of the results were consistent with the assumptions of neither the ILH nor the TFA. For instance, the results indicated that the second task was superior to other tasks in terms of its impact on vocabulary learning, although it had been given the same index as those of the first and fourth tasks by the ILH and it had been given a lower index compared to the third task by the TFA. Other results were consistent with both the ILH and TFA. An example is the superiority of the second task over the third task, which was given a higher index by the ILH, and the superiority of the third task over the first and fourth tasks, which were given a higher index by the TFA. However, further results were consistent with the TFA, but not the ILH. For instance, the third task, which was given a higher index by the TFA, was found to be more effective than the first and fourth tasks with lower indices. The same task was given the lowest index by the ILH and did not turn out to be more effective than the other tasks. Hu and Nassaji concluded that the TFA has a better predictability and attributed the results to the inadequate weighting of the ILH stating that the three factors might have different weights.

Jafari Gohar, Rahmanian, and Soleimani (2018) also compared the ILH with the TFA. The tasks included sentence-writing, composition-writing, and reading plus answering comprehension questions. The first task was given the indices of 3 and 7 by the ILH and TFA respectively. The second task was given the index of 3 by the ILH and the index of 9 by the TFA. Finally, the third task indices given by the ILH and TFA were 1 and 3 respectively. In this study, two sets of statistical analyses were carried out, one set tested the target words during

the performing of the tasks and the other set tested them in pre-tests and post-tests. The first set of analyses indicated that only the first and second tasks were not significantly different from each other. The results of the second set of analyses showed that the second task was the most effective one, followed by the first task. The Third task was the least effective one. Jafari Gohar, Rahmanian, and Soleimani concluded that the TFA is not a better predictor compared with the ILH. It could explain the superiority of the composition-writing task in the change from the pre-test to the post-test scores but could not explain the during-task performance of the groups.

2.7 Summary and Research Questions

Reviewing previous studies shows conflicting evidence regarding the ILH assumptions ranging from full support of the hypothesis (e.g., Huang, Willson & Eslami, 2012; Nassaji & Hu, 2012) to partial support (e.g., Kim, 2011; Zou, 2012; Zou, 2017) or even entire rejection of the hypothesis (e.g., Folse, 2006; Martinez-Fernandez, 2008). Although the ILH is a well-researched framework in the area of vocabulary learning, it seems that some questions remain unanswered about the hypothesis assumptions. Studies that found only partial support or rejected the hypothesis attributed their findings to issues such as the difference between initial and long-term retention of words (Kim, 2011), the level of awareness induced by tasks (Martinez-Fernandez, 2008), the effect of the number of retrievals (Folse, 2006), or contextual clues and word frequency (Bao, 2015). The ILH has also been compared with another framework (TFA) and conflicting evidence was found regarding which one is better for designing and analysing vocabulary learning tasks (Hu & Nassaji, 2016; Jafari Gohar, Rahmanian, & Soleimani, 2018).

The fact is that Laufer and Hulstijn (2001), who proposed the hypothesis, called for research to define the involvement factors precisely and to examine the contribution of each factor to vocabulary learning. However, most of the studies mentioned above were carried out using the original weight and degrees given to the factors, which may be the reason behind the conflicting evidence. Until recently, only a few studies focused on the effectiveness of the various factors instead of testing the assumptions. It means the weight of each degree of each involvement factor is not still clear. Studies by Zou (2012) and Zou (2017) proposed new degrees for the search and evaluation factors, which is a step forward in clarifying the contribution of different degrees of these factors to vocabulary learning. Bao's (2015) study questions the weighting of moderate and strong degrees of need. However, no new degree has

been proposed. More evidence is still needed to clarify the contribution of each degree of each involvement factor to vocabulary learning.

Since the ILH is a simple and valuable framework for designing and analysing vocabulary-learning tasks, the present study aims to focus on developing the hypothesis. There are several reasons for preferring this framework over the other mentioned frameworks. First, unlike Nation's approach, the ILH allows for quantification of the processes involved in learning, which makes it an easy way of differentiating between tasks and comparing them. Second, as mentioned in Chapter 1, the strength of the ILH is its simplicity. This approach with its three factors is understandable for teachers with less specialized knowledge of pedagogy as well as for autonomous learners, and it can be an easy and quick way of designing and analysing vocabulary-learning tasks. Although three factors cannot account for all parameters that affect vocabulary learning, even the TFA with its 18 criteria, as Nation and Webb (2011, p. 15) acknowledged, has not taken into account all the influential factors. Although the inclusion of more factors makes an approach more comprehensive for analysing tasks, it also makes it a complicated way of designing tasks. In addition, the study by Jafari Gohar, Rahmanian, and Soleimani (2018) indicated that the TFA is not a better predictor compared with the ILH. Hence, by developing the ILH, it may be possible to have an approach that can quantify learning processes and its simplicity has the advantage of being easy to understand and use. To this end, the present study investigates the following questions:

1- What is the contribution of each type of evaluation to vocabulary learning?

    1a: Based on the answer to the first research question, is it possible to give more degrees to the evaluation factor?

2- What is the contribution of search to vocabulary learning? (alone and in combination with different types of evaluation)

    2a: Based on the answers to the first and second research questions, it is possible to detect the impact of the distribution of the factors on vocabulary learning?

3- What is the contribution of each degree of need to vocabulary learning?

Chapter 3: Methodology

3.1 The Study

This study comprised two parts: the pilot study and the main study. The pilot study was conducted to test the suitability of the vocabulary teaching materials (two lessons) and tests (a homogeneity test, a vocabulary pre-test, and three post-tests). This chapter is presented in three major parts. In the first part, details about the designing of the teaching materials and tests are provided, accompanied by ethical considerations for using teaching materials and tests. In the second part, the pilot study, including all procedures for recruiting the participants and carrying out the procedures are elaborated. In the third part, information about the main study is given.

3.2 Designing Teaching Materials and Tests

3.2.1 Designing Teaching Materials based on the Involvement Load Hypothesis

Teaching materials had to be designed in a way to induce the involvement factors with different degrees for different groups of participants. The Involvement Load Hypothesis (ILH) rating system adopted for materials design was the combination of the original one proposed by Laufer and Hulstijn (2001), which gives three degrees to need, two degrees to search, and three degrees to evaluation and the one proposed by Nation and Webb (2011, p. 4), which gives a higher involvement load to productive retrieval and a lower involvement load to receptive retrieval. The following paragraphs provide details of the combining of the two rating systems for materials design.

As Laufer and Hulstijn (2001) elaborate, if teachers ask learners to understand or use words or if understanding words is necessary for successful completion of tasks, need is externally imposed and appears in a moderate degree, which is shown by (+). However, if learners themselves decide to express a concept for which they lack vocabulary knowledge and decide to learn a specific word to express the concept, need is self-imposed and appears in a strong degree, which is shown by (++). It is also possible that learners do not need to understand or use words, or words are not needed for task completion, or teachers do not ask learners to learn words. In such circumstances, need is absent and is shown by (-). Considering these definitions, in this study, two groups of words were included in order to investigate the contribution of different degrees of need to vocabulary learning. The first group, called the target words, were those for which knowing the meanings was necessary for task completion.

Thus, based on the ILH, the need factor for the target words was moderate. The second group, called the non-essential words, are those for which knowing the meanings was not necessary for completing the task. Learning these words depended only on the participants' motivation. Therefore, based on the ILH, the need factor for the non-essential words was strong if the participants were motivated to learn them.

Regarding the search factor, Laufer and Hulstijn (2001) state that, if the meanings of words are given, search is absent and is shown by (-). Otherwise, learners must look up the words in dictionaries, so that search appears and is shown by (+). Accordingly, in the present study, while for some groups of learners the meanings of words were provided in the form of glosses, other groups had to look up the words in dictionaries.

Evaluation, as Laufer and Hulstijn (2001) state, involves a comparison of a given word with other words or a specific meaning of a words with its other meanings. It can also involve providing words in original contexts. The former condition induces moderate evaluation, shown by (+) and the latter strong evaluation, shown by (++). If no comparison is made and words are not used in original contexts, evaluation is absent and is shown by (-). When Nation and Webb's rating system, which differentiated between receptive and productive retrieval, was applied to the definition of evaluation provided by Laufer and Hulstijn (2001), giving more degrees to evaluation could be possible (as mentioned in 1.5). Considering the new degrees, making comparisons between words does not necessarily induce moderate evaluation. Depending on how the comparison is made, either a moderate or a strong degree of evaluation can be induced. When comparisons are made between words, if the meanings of words are retrieved from memory (receptive retrieval), a moderate degree of evaluation is induced. However, if the forms of words are retrieved (productive retrieval), a strong degree of evaluation is induced. Providing words in original contexts can be in the form of writing a separate sentence for each target word (sentence-writing) or using all the target words in a paragraph/s (composition-writing). Based on the ILH, both activities induce a strong degree of evaluation shown by (++).

A detailed display of the involvement factors induced by the target words presented in the teaching materials is provided in Table 3.1. As mentioned in 1.5, for investigating the involvement factors in the present study, when evaluation is considered, it is differentiated by its different types, including receptive retrieval, productive retrieval, sentence-writing, and composition-writing. However, when need and search are considered, they are differentiated by their degrees of prominence. In the table, the involvement index for each group, which is the sum of the degrees of the factors, is also provided. The table shows ten groups of learners

because that was the number to be included in the main study. However, in the pilot study, only Groups 2 and 8 participated. The reasons for choosing these groups for the pilot study are discussed in 3.3.2.2. As presented in Table 3.1, the nine experimental groups experienced different combinations of the involvement factors with different degrees. For the control group, who did not receive the lessons, the presence/absence of the involvement factors and their degrees was not predictable. Each participant in this group might decide to learn the words after the pre-test as a result of seeing them in the test (+ need), to learn them outside the research context (++ need), or to ignore them (- need). They might also decide to evaluate the words (+ or ++ evaluation) or ignore them (- evaluation). Table 3.2 presents the involvement load induced by the non-essential words presented in the teaching materials.

It is worth adding that in Table 3.1, the involvement indices are shown based on the participants' groups. The reason is that in each group, the participants encountered all the target words under the same condition set by the tasks. In contrast, for the non-essential words, the involvement indices were not determined by the tasks. In fact, each participant's decision to ignore or search for the words could affect the degree of involvement factors. Therefore, for the non-essential words, the involvement indices are shown in Table 3.2 based on the words, not based on the groups.

Table 3.1

The involvement factors induced by the target words for groups in the main study

Groups	The induced involvement loads			Involvement index
Experimental Groups				
1	Need +	Search -	Evaluation (receptive retrieval) +	2
2	Need +	Search -	Evaluation (productive retrieval) ++	3
3	Need +	Search +	Evaluation (receptive retrieval) +	3
4	Need +	Search +	Evaluation (productive retrieval) ++	4
5	Need +	Search -	Evaluation (sentence-writing) ++	3
6	Need +	Search -	Evaluation (composition-writing) ++	3
7	Need +	Search +	Evaluation (sentence-writing) ++	4
8	Need +	Search +	Evaluation (composition-writing) ++	4
9	Need +	Search +	Evaluation -	2
Control Group				
10	Need -, +, or ++	Search - or +	Evaluation -, +, or ++	0-5

Table 3.2

The involvement factors induced by the non-essential words for groups in the main study

	The participants' decision	Need	Search	Evaluation	Involvement index
The non-essential word	Ignore the word	-	-	-	-
	Search the meaning of word	++	+	-, or +, or ++	3-5

Section 3.2.3.1 provides more detailed explanations about how the teaching materials induce each involvement factor with a specific degree of prominence. After planning the teaching materials, the target and non-essential words presented in the teaching materials had to be chosen.

3.2.2 Selecting the Target and Non-Essential Words

For designing vocabulary-focused materials, the target and non-essential words had to be chosen. Several factors that could affect learning, including frequency (Schmitt, 2010 b, p. 63), part of speech (Schmitt, 2010 b, p. 160), and concrete/abstract characteristics (De Groot & Keijzer, 2000), were controlled. The importance of considering frequency of words for research is emphasized by Schmitt (2010 b, p. 14), who points out that learners learn higher frequency vocabulary earlier than lower frequency vocabulary. Thus, the frequency of words should be matched with participants' level of proficiency. Considering that the participants of the present study were at the intermediate level of proficiency and had already learned the highest frequency words, these words could not be included in the list of the target words. According to Milton (2009, p. 186), learners at the intermediate level know words at the 1000 and 2000 frequency levels. Thus, for the present study, the target and non-essential words had to be chosen from words at the 3000 and 4000 frequency levels, on the assumption that they were unlikely to be known for the participants. The VocabProfiler (BNC-COCA-25) provided by the Compleat Lexical Tutor (v. 8.3) was used to obtain the frequency of words. The VocabProfiler has been designed based on Laufer and Nation' (1995) Lexical Frequency Profile and is used for research and teaching purposes. The VocabProfiler determines words' frequencies up to 25th thousand frequency list by analysing research corpora, namely the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). The selected target and non-essential words were typed into the submit window of the VocabProfiler. The program's output showed the frequency of words by highlighting different

frequency bands in different colours. All of the teaching and testing materials were checked for frequency of words in the same manner.

In addition to frequency, the words' part of speech and concrete/ abstract characteristics were considered. Only abstract words that were verbs, nouns, and adjectives were selected. In total, the target words were 10 nouns, 10 verbs, and 10 adjectives and the non-essential words were 4 nouns, 4 verbs, and 4 adjectives as presented in Table 3.3.

Table 3.3

Target and non-essential words Selected for the pilot study

	Lesson 1					**Lesson 2**			
	Target words	Frequency bands	Part of speech	Concrete/ Abstract		**Target words**	Frequency bands	Part of speech	Concrete/ Abstract
1	Phenomena	3000	Noun	Abstract	1	Majority	3000	Noun	Abstract
2	Emission	3000	Noun	Abstract	2	Impact	3000	Noun	Abstract
3	Climate	3000	Noun	Abstract	3	Isolation	3000	Noun	Abstract
4	Temperature	3000	Noun	Abstract	4	Addiction	3000	Noun	Abstract
5	Agriculture	3000	Noun	Abstract	5	Deficiency	4000	Noun	Abstract
6	Gradual	3000	Adjective	Abstract	6	Fundamental	3000	Adjective	Abstract
7	Global	3000	Adjective	Abstract	7	Innovative	4000	Adjective	Abstract
8	Exceptional	3000	Adjective	Abstract	8	Beneficial	4000	Adjective	Abstract
9	Vital	3000	Adjective	Abstract	9	Diverse	3000	Adjective	Abstract
10	Devastating	3000	Adjective	Abstract	10	Precious	4000	Adjective	Abstract
11	Alert	3000	Verb	Abstract	11	Devote	3000	Verb	Abstract
12	Originate	4000	Verb	Abstract	12	Restrict	3000	Verb	Abstract
13	Conserve	3000	Verb	Abstract	13	Permit	3000	Verb	Abstract
14	Confront	3000	Verb	Abstract	14	Enforce	3000	Verb	Abstract
15	Persuade	3000	Verb	Abstract	15	Decrease	3000	Verb	Abstract
	Non-essential words					**Non-essential words**			
1	Destruction	3000	Noun	Abstract	1	Intimacy	3000	Noun	Abstract
2	Expansion	3000	Noun	Abstract	2	Aim	3000	Noun	Abstract
3	Rescue	3000	Verb	Abstract	3	Irritating	4000	Adjective	Abstract
4	Endeavour	4000	Verb	Abstract	4	Illogical	3000	Adjective	Abstract
5	Blazing	4000	Adjective	Abstract	5	Participate	3000	Verb	Abstract
6	Numerous	3000	Adjective	Abstract	6	Acquire	3000	Verb	Abstract

In the teaching materials and tests, apart from the target and non-essential words, all the other words belonged to the 1000 and 2000 frequency levels, so that they would be known to the participants. An exception was made for several words such as *paragraph*, *text* and *factor*, which are lower in frequency but are also familiar because they are loan words in Farsi. The word *false* is another example that was not in the acceptable range but is well-known to Iranian learners due to its frequent use in language learning materials.

3.2.3 Lessons

After the selection of the target and non-essential words, they had to be included in learning materials (lessons), which are explained in the following sections in two separate parts. The first part indicates how each involvement factor is induced by the lessons with a specific degree of prominence. The second part elaborates the procedures for teaching each part of the materials.

3.2.3.1 Inducing the Involvement Factors

The participants had to receive two lessons designed to be similar in format (Appendix 1). Before explaining the lessons in detail, it is worth adding that in the present study, as discussed before in 2.2, the materials for presenting new words (including Sections 1 and 2 of each lesson) were comprehension-based input tasks. The rest of materials used for inducing different types of evaluation (Sections 3 of each lesson) were word-focused activities. Each of the mentioned sections and the reasons behind using them are elaborated in the following paragraphs.

Each lesson presented 15 target words and 6 non-essential words. The first lesson with the topic of 'Extreme Weather' discussed the reasons behind climate change. The topic of the second lesson was 'The Internet' and focused on the advantages and disadvantages of using the internet for children. In both lessons the contexts in which the target and non-essential words appeared were mostly unambiguous and facilitated the learning of only one meaning of each word, as intended by the tasks.

The content of each lesson was divided into three general sections. Section 1 of both lessons, including comprehension-based input tasks, was divided into Parts A, B, C and D (Appendix 1, pp. 155-157 and pp. 162-164) and was designed to provide the initial exposure to the target and non-essential words. The reason behind choosing this kind of task was that in this study, the involvement factors and their degrees of prominence had to be controlled strictly. For instance, if the participants had engaged in uncontrolled production using the target words

in the first section, the tasks would have induced strong evaluation that was not desirable for some of the groups, such as Groups 1 and 3 described in Table 3.1. Comprehension-based input tasks, which on the one hand, don't require production and on the other hand, require learners to show they have processed the input, could have been appropriate options for this research purpose. Except for Part A, which was designed as a warm-up activity, this section was designed to induce moderate need (+) for all groups and search (+) for some of them as mentioned in Table 3.1. Moderate need was induced because for completing this section, knowing the 15 target words in each lesson was necessary for task completion. For instance, in Lesson 1 (Appendix 1, p. 155), the participants needed to know the meanings of the target words referring to specific weather conditions to be able to recognize which of them have happened in their country and locate the affected areas on the map of the country as shown in Figure 3.1.

Part B

Which of the following conditions have happened in your country?

Please work individually. One of you circle the conditions that have occurred and the other circle the conditions that have not occurred. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.

For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

1- Devastating storm	2- Exceptional hot weather	3- Strong winds	4- Fall in sea levels
5- Gradual changes in weather	6- Low temperature	7- Lack of water	8- Poor agriculture

Now, please work in pairs and look at each other's answers. Does your friend agree with you?

Please work together and write the number of each condition on the map to show where the condition has occurred (An example is number '3' on the map).

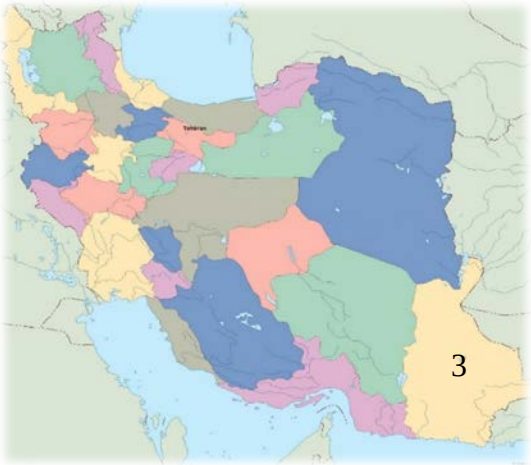


Figure 3.1

Part B of Lesson 1 used in the pilot study

In both lessons, for Section 1, the target words were provided at the bottom of the pages with or without their meanings, depending on the participants' groups. The groups for whom the search factor did not have to be induced were provided with the definitions of the target words at the bottom of the pages (as glosses). The definitions of the words were copied from the Oxford Online Intermediate Learner's Dictionary and Longman Online Dictionary of Contemporary English. In the definitions provided for the participants, low frequency words (beyond the 1000 and 2000 frequency levels) were replaced by higher frequency ones (within the 1000 and 2000 frequency levels). The groups for whom the search factor had to be induced only received the target words and the meanings were deleted from the bottom of the pages. They had to look up the meanings of words in monolingual paper or online Oxford learner's dictionaries. The participants mostly used online dictionaries on their mobile phones. Only in some cases were paper dictionaries used. A comment was also provided for them showing that looking up the meanings of words presented at the bottom of the pages might help in completing the activities. The participants were also instructed to ask for help from their teacher if they encountered any difficulty using monolingual dictionaries or understanding the meanings of words. The reason for choosing monolingual dictionaries was to provide consistency among the groups. Because the groups who received the meanings of words saw their English definitions, the other groups who looked up the meanings had to encounter definitions written only in English. Section 1 did not aim to induce evaluation.

Section 2, which also included a comprehension-based input task, was designed to present the non-essential words (Appendix 2, p. 157 and p. 164). Completing this task did not depend on knowing the meanings of the non-essential words. Therefore, the need and search factors induced in this section depended on the participants' decisions to either ignore or decide to look up the meanings of the words. A strong degree of need was induced only when the participants decided to find out the meanings of the non-essential words. The need to know the words could in turn result in looking them up (search +). However, for the non-essential words that were ignored by the participants, the need factor was absent. In the absence of need, no search was induced. Unlike Section 1, in which only the participants in the search groups were allowed to use monolingual dictionaries, all groups were permitted to do so in Section 2 because motivated participants might decide to search for the meanings of the non-essential words. Section 2 did not aim to induce evaluation. Figure 3.2 shows Section 2 of Lesson 1.

Section 2

Please read the following text individually and list the factors that cause or do not cause heating of the Earth in the below table.

‘Changes in weather are natural’ The Earth is affected by many natural events and one of them is weather change. Changes in weather have happened several times in the past and the Earth has often been warmer than now. Heating of the Earth is not caused by just one factor. One important factor during the past hundred years has been the Sun. The sun warms the Earth and a warmer planet will have more extreme weather such as storms, hot weather, and unusual rain. The earth’s weather can also change because of cloud cover. Therefore, it is not true to say that human activities such as cutting trees, using too much energy, and population growth cause heating of the Earth. The reason for changes in weather is mostly natural. Then, there is very little we can do for heating of the Earth. We may control numerous factors such as population expansion and forest destruction, but we cannot control the blazing sun. The best thing we can do is to endeavour to reduce the effects of weather events. An example is warning people before storms that can rescue people’s lives and helping victims after that.	
Factors that cause heating of the Earth	Factors that do not cause heating of the Earth

Figure 3.2

Section 2 of Lesson 1 used in the pilot study

Section 3, which included word-focused activities, was designed to induce different types of evaluation for different experimental groups, except for one of them (Group 9 in Table 3.1). In this section, the degree of evaluation induced for each group had to be controlled strictly. For groups who were meant to experience moderate evaluation (receptive memory retrieval), a gap-filling activity was designed in which the participants were given some sentences with gaps and some words at the bottom of the page. The participants had to retrieve the *meanings* of words from memory to be able to fill the gaps (Appendix 1, p. 158 and p. 165). For groups who had to experience strong evaluation, three types of activities were designed. One of them, which was called productive memory retrieval, was a gap-filling activity in which the participants had to retrieve the *forms* of words from memory to be able to fill the gaps in each sentence (Appendix 1, p. 159 and p. 166); the words were not given at the bottom of the page. The other two were sentence-writing and composition-writing in which the participants were required to provide original sentences for the target words. For sentence-writing, a separate sentence had to be written for each word. For composition-writing, a paragraph including all the target words had to be written (Appendix 1, pp. 160-161 and pp. 167-168). It was predicted that the retrieval activities would be done faster than the writing activities. Thus, an extra writing activity was designed for the retrieval groups to ensure Section 3 would take a similar amount time for all

groups. This writing activity, which was on a topic different from the lessons' topics, was meant to fill the extra time of the retrieval groups and was not used as a source of data in this research.

In each lesson, the instruction for each part showed whether the part had to be done individually or in pairs. All the instructions were also explained by teachers. Throughout both lessons, except Part A which was a warm-up activity, the pair activities were preceded by the individual ones. The aim of individual activities was to help the participants to focus on the target words and to search for their meanings if asked to do so (in Section 1), to decide about learning or ignoring the non-essential words (in Section 2), and to evaluate the words individually without being affected by another participant (in Section 3). In Section 1, after ensuring that all participants could focus on the aims, they could do pair activities, which were designed to make the lessons more interesting. A summary of the lesson structure is presented in Table 3.4.

Table 3.4

The structure of the lessons

	Type of activities	Purposes	Use of dictionary
Section 1			
Part A	Individual work	Warm up	Only allowed for the participants of the search groups
Part B	Individual work Pair work	• Inducing need for the target words for all groups • Inducing search for some groups	Only allowed for the participants of the search groups
Part C	Individual work Pair work	• Inducing need for the target words for all groups • Inducing search for some groups	Only allowed for the participants of the search groups
Part D	Individual work Pair work	• Inducing need for the target words for all groups • Inducing search for some groups	Only allowed for the participants of the search groups
Section 2			
	Individual work	Presenting the non-essential words that may or may not induce need and search	Allowed for all the groups
Section 3			
	Individual work	Inducing evaluation for the target words	Not allowed

The table shows how each section/part of the lessons induced the involvement factors. The following section elaborate the teaching procedures for each section in both lessons.

3.2.3.2 Procedures for Teaching Each Part of the Material

Section 1, Part A: The learners were given a question about the topic and then, working in pairs, they wrote words or sentences related to the topic of the lesson. This part did not contain any new word and did not induce any of the involvement factors. Its aim was merely to activate schemata and prepare the participants for the content of the lessons through brainstorming. At the end, the teacher asked a few pairs to read their words/sentences to the class.

Section 1, Part B: The participants were asked to do an activity for which they needed to know the meanings of some unknown words (the target words). In this part, while some of the groups were given the meanings of words, others had to search for them. The groups who received the meanings of words were instructed to look at the bottom of the page if they needed to know a word. The search groups were instructed to use monolingual dictionaries if they needed to know a word.

In the beginning of Part B, the participants were asked to work individually. In Lesson 1, while one of the partners in a pair circled the conditions that had occurred before, the other circled the conditions that had not occurred. In Lesson 2, each of the participants wrote the numbers of the sentences they agreed with. At the end of Section B, the participants of each pair had to cooperate to complete the task. In Lesson 1, the participants were required to look at each other's answers and, working together, write the number of each condition on the map to show where each condition had occurred before. In Lesson 2, each participant had to tell the partner the numbers of the sentences they agreed with to help them guess their view. Thus, this section did not allow for language production. The output was writing numbers on the map or in the table.

Section 1, Part C: The same as Part B, this part contained some unknown words. First, the participants worked individually and expressed their ideas about the reasons for changes in weather (in Lesson 1) and the advantages and disadvantages of using the internet for children (in Lesson 2) by choosing yes, no or not sure. After completing this part, they were invited to work in pairs and compare their answers. If their partners did not agree with them, they had to go around the class and find someone who agreed with them. Then, in Lesson 1, each pair were asked to categorize the reasons for changes in weather based on their importance by putting the number of each reason in one of the three categories of very important, important, and not important. In Lesson 2, each pair had to categorize the advantages and disadvantages of using the internet based on their frequency of occurrence by putting the number of each in one of the

three categories of usually happens, may or may not happen, and never happens. The same as Part B, the output was writing numbers.

Section 1, Part D: This part also contained some unknown words. First, the participants worked individually. One of them had to read some sentences containing the target words and expressed his/her ideas by choosing Agree, Disagree, or Not sure. The other one had to read the same sentences, guess his/her friend's ideas, and choose Agree, Disagree, or Not sure to reflect the partners' ideas. Then, in the pair activity, pairs had to show how many right guesses they had made. Because partners of each pair had already done Part C together, they knew each other's opinion about the subject of the lesson. The output was numbers.

Section 2

This section involved reading a short passage designed to present the non-essential words. The paragraphs of both lessons had nearly the same length, 194 and 193 words. The number of non-essential words included in each of the paragraphs was the same (6 words). The rest of the words of both paragraphs were within the 1000 and 2000 word frequency lists. Thus, Section 2 was quite similar across the groups. Each non-essential word appeared once in the text in order to have the same number of exposures as the target words. No target word was provided in this section. The participants had to read the passage related to the topic of the lesson individually and fill in a table using information from the text. No pair activity was permitted in this section. Searching for or knowing the meanings of the non-essential words could not affect task completion, which was filling the table.

In this section, the participants of all the experimental groups were also instructed to circle the words in the paragraph they had looked up in a dictionary while doing the tasks. This was done to keep track of the non-essential words the participants decided to learn based on their own motivation. For these words the need factor was strong (++). However, for the participants who had not looked up the meanings of the non-essential words, any score on the post-tests related to the non-essential words was not taken into account. The reason was that if the participants had decided to learn the words after the lessons or outside the context of the study, determining the degree of need could be impossible. If learning the words was the result of taking the tests, the need factor would be moderate. If learning happened outside the research context, the need factor could appear in either a strong or a moderate degree. Due to this uncertainty, no score was given to the correct answers in the post-tests if the participants had

not circled the words in Section 2 showing that they had not searched their meanings during task completion.

Section 3

Unlike the previous ones, this section did not include input-based tasks and presented only word-focused activities. No pair activity and no dictionary use were allowed in this section to prevent any degree of search from being induced. Before doing the activity, the participants were asked to put Sections 1 and 2 aside and not to look at them; they were also not to use their dictionaries. The participants who completed the gap-filling activity had to express their ideas about the sentences by circling yes, no or not sure. For the groups who experienced receptive memory retrieval, for some of the sentences, more than one word could make sense in the context. The learners' answers were acceptable as long as the chosen words were appropriate for the given contexts.

During the evaluation process, the participants could receive help only from their teachers. The main aim of this activity for all the groups was to provide a condition for the participants to evaluate the new words. During the evaluation process, the participants had to be monitored and encouraged to make decisions about the words that could fit the contexts or to use the words in original sentences. The outcome of this activity was the sentences with filled gaps, separate sentences written for each target word, or short paragraphs written using the target words. After completing the activity, the participants had to be instructed to look at Sections 1 and 2 and check their performance. The teacher was able to comment on the participants' performance if they wished.

3.2.4 Tests

Tests encompassed a homogeneity test, a vocabulary pre-test and three vocabulary post-tests. In all tests, the verbs that were presented as target and non-essential words in the teaching materials appeared in their base forms and all the nouns in singular forms. The contexts in which the target and non-essential words were provided were mostly unambiguous. In addition, tests only measured those meanings of the words that had been presented in the teaching materials.

3.2.4.1 Homogeneity Test

Based on the researcher's previous experience and familiarity with the research context, it had been predicted that administering a comprehensive test to assess the homogeneity of the groups was not possible due to practical issues such as time limitation. For this reason, before the start of the treatment, the X-Lex test (Appendix 15) designed by Milton (2009), which could be done in less than 15 minutes, was administered to check the participants' level of proficiency and the groups' homogeneity. The X-Lex test is a passive recognition test that estimates language learners' vocabulary size with a ceiling of 5000 words. Use of the X-Lex test to check the participants' level of proficiency was based on the assumption stated by Milton (2009, pp. 171-172) that vocabulary knowledge is a good indicator of language proficiency and vocabulary size can be linked to the Common European Framework (CEFR) levels. Milton (2013, pp. 71-72) also states that because of the close connection between vocabulary size and language skills, vocabulary size can be a reliable measure as a placement test. The X-Lex test can be administered on computer and in a paper and pencil form. Considering that most Iranian language institutes, where the data is collected, do not have enough facility to administer the test on computers, the test was administered in a paper and pencil form. The participants were given a list of 120 words, including 20 pseudo words and 100 real words, and were asked to tick the words they know. The real words were sampled randomly from the most frequent 5000 lemmatised words in English, with 20 words from each 1000 band. The pseudo words provide the possibility of calculating over-estimations of knowledge and adjusting the scores. In the original test provided by Milton, the words are organized in six columns. One column is for the pseudo words and each of the other columns includes the words of a specific frequency level. However, in this study, considering Milton's (2009, p. 253) suggestion, the presentation of the real and pseudo words was randomized, as shown in Appendix 15. In the first two rows of the table, no pseudo words were provided to avoid demotivating the participants. The test takers were told that some of the words were real English words, but some of them that looked like real words were pseudo words and they were required to tick only the words they knew. Because the words were randomized in the X-Lex test, it made no difference for the participants to work across or down the table.

For scoring the test, for each real word ticked by a test taker 50 marks were awarded and the marks for all the real words ticked by the test taker were added up. Then, 250 marks were subtracted from the raw score for each pseudo word ticked by the test taker. The result, which was an adjusted score, was an estimate of the test taker's vocabulary size (Milton, 2009,

p. 253). The test taker's vocabulary size had to be interpreted using Table 3.5. This table is part of the original table provided by Milton (2009, p. 186). Because in the present study the participants had to be at the intermediate level of proficiency, the test takers with a vocabulary size around 2500-3250 were chosen. However, it was not possible to exclude those students whose scores did not fall in the acceptable range from their classes. These students attended all the treatments and tests, but their scores were not included in the data.

Table 3.5

Mean EFL vocabulary size scores and the CEFR (Milton, 2009, p. 186)

CERF level	Wordlist size	X-Lex
A1		<1500
A2	1000	1500-2500
B1	2000	2500-3250
B2		3250-3750
C1		3750-4500
C2		4500-5000

3.2.4.2 Vocabulary Pre-test

3.2.4.2.1 Rationale for the Vocabulary Pre-test

The vocabulary pre-test was designed considering Schmitt's (2010 b) view about measuring learners' knowledge of words. According to him, asking learners, especially those at lower levels of proficiency, to show their knowledge of the words by yes/no self-reports results in false reporting that they know the meanings of words. They in fact may be confusing the words with other words they know with a similar form (p. 159). Thus, in the present study, the participants' initial knowledge of the target and non-essential words was measured directly by asking them to write the L1 meanings of the L2 words. It seems that measuring the participants' knowledge of words based on providing their L1 meanings is more valid compared to yes/no self-report.

It may be argued that this kind of pre-test exposes the target and non-essential words to the participants in advance so that they may pay more attention to them during the treatment. However, as will be explained in the next section, the pre-test included 60 words, nearly one third of which were distractors. Due to the large number of words presented to the participants, it seems unlikely that they clearly remembered most of them during the treatment and paid more attention to them. It is worth adding that even if the participants remembered the target

words and paid more attention to them, the results of the study could remain unaffected because the objective of the treatment was drawing the learners' attention to the target words in order to learn their meanings. Only learning the non-essential words, which was a decision that had to be made by the participants, might have been affected. However, as will be explained in 4.3.4, careful study of the participants' papers showed that many of the participants ignored the non-essential words in both lessons, indicating that they did not remember them.

3.2.4.2.2 Vocabulary Pre-test Specifications

In the beginning, a vocabulary pre-test had to be given to the participants to measure their initial knowledge of the target and non-essential words (Appendix 2). In total, 30 target words, 12 non-essential words, and 18 distractors were included in the vocabulary pre-test. The distractors were mostly taken from the text of the teaching materials and tests of the present study and belonged to the 1000 and 2000 frequency lists in order to be familiar for the participants. In the pre-test, the participants were required to show their knowledge of the meanings of words by providing their L1 translations. The words were organized into two columns with a space in front of them for writing their meanings. The reason for providing distractors was to motivate the participants to complete the test. Because nearly two third of the words in the pre-test had been predicted to be unknown for the participants, they might not have given enough attention to the words, thinking that they knew none of them. The distractors with high frequency would help them to feel they could complete the test by paying more attention to the items. The test started with two high frequency words so the learners might be motivated to continue. The distractors were distributed through the list in a way that after one or several unknown words the participants would see a familiar word (a word with a high frequency).

3.2.4.3 Vocabulary Post-tests

3.2.4.3.1 Rationales for the Vocabulary Post-tests

For designing and administering the post-tests, several issues were considered. Nation and Webb (2011, p. 279) recommend that more than one test be administered for the same words. Therefore, first, three types of tests (productive, translation, and matching) were chosen. To justify the tests' order of administration, the notion of degrees of vocabulary knowledge (Laufer, Elder, Hill, & Congdon, 2004, pp. 206-207; Laufer & Goldstein, 2004, pp. 405-406)

was taken into account. The notion defines degrees of knowledge of meanings along two dichotomies:

- Retrieving the L2 form for a given meaning, or active knowledge as compared to retrieving the meaning for a given L2 form, or passive knowledge
- Recalling the forms or meanings of words as compared to recognising forms and meanings in a set of options

In this study, the productive test was designed to measure the participants' ability to supply the forms of the target L2 words by looking at their definitions. This test can be categorized as an active recall test that, according to Laufer et al. (2004) and Laufer and Goldstein (2004), is the most difficult one. The translation test was designed to measure the participants' ability to retrieve the meanings of words by looking at their L2 forms and was a type of passive recall test. The matching test that measured the participants' ability to recognize the L2 definitions of the L2 words in a set of options was a type of passive recognition test and was the easiest one. Based on Laufer et al.'s (2004) and Laufer and Goldstein's (2004) statements that the tests should start with the most difficult one, in the present study, the productive test was administered first. Then, the translation test and the matching test followed.

3.2.4.3.2 Specifications for the Productive test

For the productive test (Appendix 3), a list of L2 definitions was provided and the participants were instructed to write an English word for each definition in front of it. The first letter of each word was provided to avoid eliciting words other than the target and non-essential words with similar meanings. For writing the definitions, the Oxford Online Intermediate Learner's Dictionary and Longman Online Dictionary of Contemporary English were consulted. The definitions had simple structures and were short and clear. The words in the definitions were within the 1000 and 2000 frequency levels to be familiar to learners at the intermediate level of proficiency. For each correct English word written by the participants, one point was given and for each word that was not one of the target and non-essential words and for no answer, no point was assigned. Partly correct answers received 0.50 point. For instance, a participant who wrote *blazy* instead of *blazing* received half a point. This method of scoring for taking into account partly correct answers by giving fewer points to them had been used in previous studies (e.g., Cohen & Aphek, 1980; Huang & Lin, 2014; Pulido, 2009;). After the test, the papers were collected and the participants were ready to receive the papers of the next test.

3.2.4.3.3 Specifications for the Translation Test

For the translation test (Appendix 4), a list of the target and non-essential words was given to the participants and they were asked to provide their L1 meanings. The participants were instructed to provide only one and the best L1 meaning for each L2 word. Each correct response had one point and a wrong or no response received no point. Partly correct answers received 0.50 point. After the translation test papers were collected, the participants received the matching test papers.

3.2.4.3.4 Specifications for the Matching Test

For the matching test (Appendix 5), the target and non-essential words were divided into groups of three or four. In front of each group, six or eight definitions were provided. This decision was based on Carr's (2011) view that in a matching test, the number of definitions should be more than the words (pp. 30 and 95). The participants were instructed to match each word on the left with only one definition on the right. The definitions were arranged by length and there was no pattern for the responses. For writing the definitions, the Oxford Online Intermediate Learner's Dictionary and Longman Online Dictionary of Contemporary English were consulted. The three or four words presented together shared the same part of speech because otherwise the participants might identify the correct answers based on their knowledge of the word class. The wrong definitions defined words with the same part of speech as the target and non-essential words. Wrong definitions had some words in common with correct definitions and mostly had similar lengths. The definitions were kept clear and short. Except for the target words, all the words belonged to high frequency levels. Each correct match had one point. Each wrong match or no response had to receive no point.

3.2.5 Ethical Considerations

Before carrying out the pilot and main studies, Ethics approval was obtained from the University of Auckland Human Participants Ethics Committee (UAHPEC) for the pilot and main studies separately (Appendices 6 and 7). The following general measures were taken to comply with ethical requirements for both the pilot and main studies.

- All students were approached by teachers of language classes.

- Heads of language institutes, teachers, and all students were informed of the details of the study (i.e., the purpose, procedures for data collection, data storage, and future use of data) and

terms and conditions for participation (i.e., voluntary of participation, confidentiality of identity, rewards for participation, right to withdraw from participation, and right to withdraw personal data within a time limit) through a Participant Information Sheet (PIS) before their participation (Appendices 8, 9, 10,).

- Heads of language institutes, teachers, and all students signed a Consent Form (CF), in which all the terms and conditions for participation were clearly explained (Appendices 11, 12, 13).

- Raters of tests signed a Confidentiality Agreement Form (Appendix 14).

3.3 Pilot Study

The context of this study was Iran where English is a foreign language and is taught in language institutes using Communicative Language Teaching (CLT). The pilot study had to be done with two groups of learners. After obtaining the consent of the head of a private language institute in Tehran/Iran, two teachers were introduced to the research aims and procedures. All teaching and testing materials of the present study were designed for learners at the intermediate level of proficiency. Thus, the teachers were chosen because they were teaching classes of learners which were considered to be at that level, based on the institute's assessment (explained in the following section). The teachers gave consent and checked the suitability of the two lessons and two tests for their students. Then, in each of the two classes, the PISs and the CFs were distributed among students by their teachers. The students' questions about the research were answered by the researcher and all of them gave consent to participate in the study.

3.3.1 Participants

The participants were adults (around the age of 18) and their mother tongue was Farsi. The course book used in the language institute was the second edition of the American English File (AEF), which is one of the most popular books for English teaching in Iran. In the language institute, at the end of each semester, all learners were required to take a written test provided by the American English File (AEF test) for their level of proficiency. Gaining 70 out of 100 was necessary for passing the semester test. The tests included seven parts, namely vocabulary, grammar, pronunciation, listening, speaking, reading, and writing. Each of Grammar and vocabulary had the highest mark of 20. Reading and speaking had 15 marks each. The lowest mark of 10 was given to each of pronunciation, writing, and listening. All students of the chosen

classes had been studying English in the same language institute for several semesters and had passed the AEF test to be placed in the intermediate classes. The students of both groups obtained good marks in the previous semester's AEF test (above 80 out of 100).

The results of the AEF test, which had been obtained from the language institute, had to be double-checked against the learners' scores on the X-Lex test. The two tests (AEF and X-Lex), could complement each other and help to identify the outliers. Both pilot groups took the X-Lex test in about 8 and 9 minutes respectively. It seemed that the test instruction was clear enough for the participants. They understood what they were supposed to do and worked accordingly. The participants' AEF scores and their scores in the X-Lex test that correspond to the CEFR levels are presented in Tables 3.6 and 3.7.

Table 3.6

The first pilot group's X-Lex and AEF scores

Participants	X-Lex test scores (vocabulary size)	CEFR levels	AEF scores (out of 100)
1	2250	A2	86
2	2450	A2	90
3	2550	B1	94
4	2350	A2	92
5	2400	A2	89
6 *	1550 *	A2 *	82 *
7	2350	A2	90
8 *	4250 *	C1 *	98 *
9	2450	A2	93
10	2700	B1	90
11	2300	A2	85
12	2300	A2	84
13	2550	B1	90
14	2400	A2	88
15 *	1000 *	A1 *	80 *
16 *	1550 *	A2 *	82 *

Table 3.7

The second pilot group's X-Lex and AEF scores

Participants	X-Lex test scores (vocabulary size)	CEFR levels	AEF scores (out of 100)
1	2400	A2	96
2	2500	B1	92
3	2400	A2	84
4	2450	A2	85
5	2600	B1	95
6 *	2250 *	A2	90 *
7	2450	A2	90
8 *	2300 *	A2	97 *
9	2650	B1	93
10 *	2250 *	A2	90 *
11	2600	B1	95

The X-Lex test results revealed two points. First, the vocabulary size of most of the participants was not within the range of 2500-3250 words that corresponds to the intermediate level (B1) but was close to it. Second, in the first group, unlike the majority of the participants whose scores fell in a range between 2000 and 3000 words, four participants' scores were below or above the range as shown by asterisks in Table 3.6. The same four participants had either the highest or the lowest scores in their AEF test. Considering the group's scores in both the X-Lex test and the AEF test, the four participants were recorded as outliers and their data were excluded from the study.

The second group's X-Lex test scores, as shown in Table 3.7, also fell within the range of 2000-3000 words. Although some of the participants, such as the ones shown by asterisks in Table 3.7, got the lowest scores in the X-Lex test, they had not done badly in their AEF test compared to other participants in their group and thus were not excluded.

Another consideration for choosing the participants was the number of languages they spoke because it was likely that some of the learners had the knowledge of a European language other than English. In the beginning of the X-Lex test, the participants had to answer a question about their language learning background. Some of the participants mentioned that they could speak Azari, which is spoken in some parts of Iran. In the last few decades, a large number of Azaris have moved to the capital city of Iran, Tehran, and nowadays, Azari is widely spoken in this city. A few participants stated that they could understand written Arabic. Although Arabic is only spoken in the south-west border of Iran, near Arabic countries, understanding written Arabic as the language of Quran has been important for highly religious people in other areas. Since speaking Azari and understanding written Arabic could not affect the participants' English vocabulary learning in this study, no one was excluded due to the mentioned language

background. Thus, in total, 12 students from the first group and all 11 students from the second group provided data for the pilot study.

3.3.2 Procedures

The procedures were carried out over five sessions for each group. In the first session, the vocabulary pre-test and the X-Lex test (homogeneity test) were administered. In the second session of data collection, the groups received Lesson 1. In the third session, three vocabulary post-tests of Lesson 1 were administered. The fourth session was for teaching Lesson 2 and in the fifth session, three vocabulary post-tests of Lesson 2 were administered. In the language institute, the sessions were held twice a week, on Wednesdays and Saturdays. The students of both groups received the lessons on Wednesdays and took the tests on Saturdays to ensure that the time between the lessons and the tests were the same for both groups. Before each session, the teachers were told how to carry out teaching and testing procedures. The goal of these procedures was collecting data about the teaching and testing materials in order to modify them based on the participants' performance and teachers' comments.

3.3.2.1 Piloting the Pre-test

The participants took the vocabulary pre-test and their papers were analysed with the aim of measuring the participants' initial knowledge of the target and non-essential words, identifying any problems in the test, and compiling an answer key for the main study. In the pre-test, the participants were required to show their knowledge of the meanings of the L2 words by providing their L1 translations. The test took nearly 12 minutes for the first group and 13 minutes for the second one. The papers were marked and analysed by two raters. The results of the vocabulary pre-test indicated that the majority of the participants knew some of the target words. Table 3.8 shows the target and non-essential words familiar to the participants, with the most familiar words on the top. As the table shows, the number of participants in the first group who knew the target and non-essential words dropped from 9 (for the second, third, and fourth most familiar words) to 4 (for the fifth familiar word), showing that the first four words shown by asterisks were highly familiar to the participants. These words had to be eliminated from the lessons and tests for the main study. Finally, to make an answer key, the raters used dictionary definitions as a reference point and reached an agreement about which of the students' answers could be acceptable for each item. This answer key containing a list of acceptable answers and dictionary definitions was used to mark the pre-test papers in the main

study (Appendix 16). However, it had to be used flexibly in evaluating the answers provided by the participants in the main study. To do this, the raters had to keep a running record of additional answers that they accepted.

Table 3.8

The target and non-essential words familiar to the participants

Target and non-essential words		The number of students in the first group who knew the word	The number of students in the second group who knew the word	Total number for each word
1	Global*	10	8	18
2	Temperature*	9	5	14
3	Alert*	9	3	12
4	Decrease*	9	1	10
5	Rescue	4	4	8
6	Originate	5	1	6
7	Participate	4	1	5
8	Aim	2	2	4
9	Destruction	2	1	3
10	Conserve	3	0	3
11	Phenomena	0	2	2
12	expansion	2	0	2
13	Blazing	1	1	2
14	Beneficial	0	1	1
15	Impact	0	1	1
16	Addiction	0	1	1
17	Isolation	0	1	1
18	Irritating	0	1	1
19	Illogical	0	1	1
20	Fundamental	0	1	1
21	Confront	1	0	1
22	Persuade	1	0	1
23	Enforce	1	0	1
24	Numerous	1	0	1

The inter-rater reliability of the pre-test was also calculated using SPSS Statistics version 25. Cronbach's Alpha for both groups' pre-tests had the same value of .997 that was high enough to show the reliability of scoring.

3.3.2.2 Piloting Teaching Materials

For the pilot study, the time limit for doing the tasks was the length of the class session. The teachers had to announce the end of each task when the participants had completed it. The completion time of each task was recorded for the groups and the average time was the basis for determining the task completion time in the main study. For all the instructional materials, the participants were asked to write their names at the top of the papers.

In the main study that involved 10 groups, the various types of evaluation would be induced by giving a different word-focused activity to each group. However, in the pilot study with two groups, only two types of evaluation could be induced. Choosing the type of evaluation for the pilot study had been based on the prediction that composition-writing might take longer than sentence-writing. Productive retrieval that required learners to provide the *forms* of the words might also take longer than receptive retrieval, which asked learners to retrieve the *meanings* of the words. Including more time-consuming activities in the pilot study and recording the time on task could help to estimate the time of the shorter activities and to ensure that in the main study, all the groups could complete their tasks in the allocated time. Thus, the two pilot groups needed to experience composition-writing and productive retrieval respectively. Then, decisions had to be made about the search factor induced for the pilot groups. Because looking up the meanings of words in dictionaries was more time-consuming than reading the glosses, the composition-writing group was asked to look up the words. The productive retrieval group was given glosses to see whether they could readily understand the definitions. The chosen groups for the pilot study were Groups 2 and 8 shown in Table 3.1. The involvement load induced for the target and non-essential words for these groups are presented in Table 3.9.

Table 3.9

Tasks for the pilot study

Groups	Section 1 (for the target words)		Section 2 (for the non-essential words)			Section 3 (for target words)	Involvement indices for the target words	Involvement indices for the non-essential words
8	Need +	Search +	Either - or ++ Need	Either - or + Search	Either – or + or ++ Evaluation	Evaluation (composition writing) ++	4	0-5
2	Need +	Search -	Either - or ++ Need	Either - or + Search	Either – or + or ++ Evaluation	Evaluation (productive retrieval) ++	3	0-5

As mentioned before, each of the lessons had three sections (Appendix 1) designed to induce the involvement factors. In Section 1 of both lessons, successful completion of the tasks was not possible without knowing the meanings of the target words. Thus, as shown in Table 3.9, need was moderate (+) for the target words for both groups. In this section, the productive retrieval group (group 2) was provided with glosses. However, the composition-writing group (group 8) had to look up the meanings of words in a dictionary. Thus, as illustrated in Table

3.9, in Section 1, search is (+) for the composition-writing group and (–) for the productive retrieval group. Section 2 was designed to present the non-essential words. Here, the need factor could have different values depending on the participants' motivation to know the words irrespective of the group they were in. In each of the groups, if a participant decided to know a word, the need factor would be strong (++) for the word and search factor would also appear (+). Otherwise, both need and search would be zero. Evaluation could have any degree depending on the participants' decisions outside the research context. Explanations regarding controlling this variable are provided in 4.3.4. In Section 3, which was designed to induce evaluation, for the composition-writing group that had to use all the newly learned words in a paragraph, a strong degree of evaluation (++) was induced. The productive retrieval group had to retrieve the forms of the words to be able to fill in the gaps in sentences; based on the ILH assumptions and Nation and Webb's (2011) view, this also induced a strong degree of evaluation (++).

Because it had been predicted that composition-writing would take longer than productive retrieval, the former group was one session ahead of the latter group so that the time taken by the first group could be recorded and be given to the second group. The sessions for the composition-writing group were held on Saturdays at 3 pm and for the productive retrieval group on the same day at 5 pm. In the pilot study, time on task for each group on each section was recorded and presented in Table 3.10. Time on task for each section was measured by recording the time the participants started the section and the time they completed it.

Table 3.10

Time on task for each group

	Group 1 (Search and composition writing)	Group 2 (No search and productive retrieval)
	Approximate completion time	
Lesson 1		
Section 1:		
Part A	5 mins	5 mins
Part B	13 mins	9 mins
Part C	11 mins	8 mins
Part D	13 mins	10 mins
Section 2	9 mins	11 mins
Section 3	20 mins	17 mins
Total time	71 mins	60 mins
Lesson 2		
Section 1:		
Part A	7 mins	8 mins
Part B	15 mins	10 mins
Part C	12 mins	9 mins
Part D	10 mins	7 mins
Section 2	8 mins	9 mins
Section 3	22 mins	18 mins
Total time	74 mins	61 mins

As presented in Table 3.10, for the composition-writing group, who had to look up the meanings of words, the lessons took longer. In addition, in each group, the time for completing each part of Section 1 varied according to the number of target words presented in each part. The time given to Part A of Lesson 2 was longer than the time spent on the same part of Lesson 1 because the participants in Lesson 2 had to write sentences rather than just words. In addition, the sections in Lesson 1 and 2 contained different numbers of the target words.

In addition to recording the time, teaching procedures were observed to ensure the consistency of treatment for the groups and to take note of any issues. The participants' complaints revealed some problems, including vagueness of some sentences and faults in the order of presenting information. The problems are explained in the following sections.

Problems in Section 1

In Part C of Lesson 1, as shown in Appendix 1 (p. 156), the participants were required to read two opposing groups of reasons for changes in weather. They could show their agreement or disagreement with each of the reasons by circling Yes, No, or Not sure. At the end, two opposite ideas, including 'Changes in weather originate from human activities' and 'Changes in weather are natural', had been presented as concluding remarks (the groups' ideas). Monitoring the

participants during task completion showed that the remarks had not been correctly placed in the task because most of the participants ignored them. It seemed that after reading the reasons and expressing ideas, the participants assumed that they had completed the task and did not proceed to the concluding remarks at the bottom of the photos. Since one of the remarks included the target word, *originate*, wrong placing of information made the task unsuccessful in inducing need for the target word. The same part in Lesson 2, in which the remarks had been presented in the beginning of the task (Appendix 1, p. 163), was successful in inducing need for the target word *beneficial*. In Lesson 2, it was observed that the participants focused on the target word included in the remark before proceeding to the opposing statements. Therefore, for the main study, the order of presenting information in Lesson 1 was changed. The remarks in Part C of Lesson 1 was placed on the top of the photo as shown in Appendix 17 (p. 194).

In Part D of Lesson 1, the participants were asked to read some sentences containing ideas about the role of people in climate change and to express their ideas by circling Agree, Disagree, or Not sure. The participants complained that three of the sentences were too vague. The first one is a negative sentence: 'It is not vital for people to use less gas and oil'. The sentence might not seem vague by itself, but, as the participants mentioned, answering a negative sentence could create confusion. It seemed that the sentence had to be more straightforward because the participants could not focus on the target word in the sentence, grasp the meaning of the negative sentence, and think about and express their ideas all at the same time. To clear up the confusion, this sentence was changed to a positive one for the main study. Another sentence that, in some of the participants' opinion, did not make much sense was 'We must confront changes in weather in our country'. It was changed to 'The world must come together to confront changes in weather'. The last sentence that needed revision, 'We can only reduce the effects', was changed to 'We can only reduce the effects of extreme weather'.

In addition to the sentences, there was a word in Part D of Lesson 1 that created confusion for some of the participants. The definition of *phenomena* was to some extent similar to the definition of *exceptional* in Part B of the same lesson. *Phenomena* was defined as 'unusual events' and *exceptional* as 'very unusual'. Since the definition of *phenomena* could be clearer, it was changed to 'events that are not fully understood'.

Problems in Section 2

This section had been designed to present the non-essential words. In Section 2, the participants were supposed to read a paragraph and fill in a table using the information presented in the paragraph (Appendix 1, p. 157 and p. 164). Knowing the meanings of the non-essential words

was not needed for filling in the table. Based on the learners' motivation to ignore or to look up the meanings of the non-essential words, the section could induce zero or strong need respectively.

It had been assumed that the learners would look at the table first to know what to look for in the paragraph and then would fill in the table. Contrary to this assumption, the participants first read the paragraph completely, looked up all the unknown words, and then, started filling in the table. The way the participants approached the task affected the degree of need. Instead of having self-imposed need (++), need was created by the task (+). Factors such as the task instruction and the place of the table could make the learners to see the task differently. As shown in Appendix 1, in Section 2 of both lessons, the tables had been placed after the paragraphs and the participants, who as intermediate-level learners were not well familiar with reading strategies, decided to read the paragraphs before filling in the tables. In addition, the instruction and the type of output could lead the participants to read the paragraphs first.

This section in both lessons was revised in a way that it could let learners decide about learning the non-essential words. The type of output was changed, the table moved to the beginning of the task, and the instruction was revised. Instead of asking the participants to read the text and fill in the table provided below the paragraph, they were asked to read the table and tick the options provided in the table above the paragraph (Appendix 17, p. 195 and p. 202).

Problems in Sections 3

The gap-filling activity of this section was piloted based on three main issues explained by Read (2012, pp. 309-310). The first issue about piloting gap-filling tasks is whether the participants can provide acceptable answers based on contextual clues in the sentences or not. The second issue is that if the sentences are intended to elicit one target word, they need to be written in a way to rule out other options. The pilot study revealed that these two issues had not been considered carefully for designing the gap-filling task. During the pilot study, the participants had stated that some of the sentences had been ambiguous. In addition, although the participants had experienced productive retrieval in which the first letter of each target word had been provided, they had said if they had not had the first letters of some of the words, they would have written another word. Because in the main study some of the groups would experience receptive retrieval, in which they would not have the first letters of the words, this problem had to be solved.

The ambiguous sentences and the words that could fit into more than one sentence, based on the participants' claims, were reconsidered after the pilot study. Contrary to the participants' claims, it was found out that most of these words could not match more than one sentence because they were not grammatically correct or did not make adequate sense in the context. It seemed that the participants who were struggling to remember and evaluate a number of words they had just encountered needed clearer sentences. Hence, the sentences had to be revised in a way that grammar could not affect the options and the target words could make clear senses in the view of an intermediate learner.

The sentences were reviewed and a list of the target words that could make sense in each sentence irrespective of grammar was written (the second left column of Tables 3.11 and 3.12). Then, the sentences were revised to limit the number of words that could match each context in the absence of the first letters (the third left column of Tables 3.11 and 3.12). As mentioned in Table 3.8, the words *global*, *temperature*, and *alert* in Lesson 1 and the word *decrease* in Lesson 2 were deleted from the study due to being familiar to most of the participants. The word *fundamental* was also eliminated from Lesson 2 for a reason that is explained in the following paragraphs. Thus, the sentences containing these words were also eliminated from Section 3.

Table 3.11

Revisions of Section 3 of Lesson 1

Sentences for the pilot study	target words that could match each context	Sentences after revisions	Acceptable answers
1-There are many reasons for climate change and human activity is just one of them.	Climate Exceptional Gradual Devastating	The Earth's climate is changing and human activity is the main reason for it.	Climate
2-Warmer weather and strong winds are not new phenomena in the Earth's history. In the past, the Earth was warmer than now.	Phenomena Climate	Extreme weather phenomena can endanger people's lives.	Phenomena
3-Changes in weather are gradual, so there is little danger in the near future.	Gradual Exceptional Phenomena Devastating	Changes in weather are gradual, so people may not become aware of them in their daily lives.	Gradual
4-Sometimes, weather events are devastating and normal people cannot do much to stop these effects.	Devastating Exceptional Phenomena Gradual	Sometimes, weather events are devastating and it costs a lot to repair the effects.	Devastating Exceptional
5- One reason for exceptional storms is air pollution.	Exceptional Devastating	One reason for exceptional hot weather in some countries is air pollution.	Exceptional Devastating

As shown in Table 3.11, although revising sentence 4 containing the word *devastating* limited the number of words that could match the sentence, the word *exceptional* could still make sense in the context. Because *devastating* could make sense in sentence 5 containing e*xceptional* as well, both words were acceptable for sentences 4 and 5. Revising sentence 3 containing the word *gradual* also limited the number of options for the sentence. The word *phenomena* might still make sense in the context, but the word *gradual* could not fit sentence 2 containing *phenomena*. Since the words could not be used interchangeably, only one of them was acceptable for each specific context as shown in Table 3.11. Similar revisions were made for Section 3 of Lesson 2 as presented in Table 3.12.

Table 3.12

Revisions of Section 3 of Lesson 2

Sentences for the pilot study	target words that could match each context	Sentences after revisions	Acceptable answers
1-Social isolation happens when people spend their time on the internet and do not meet family and friends.	Isolation Deficiency	Isolation from the society happens when people spend their time on the internet and do not meet family and friends.	Isolation
2- Using the internet can restrict the time people spend for daily activities.	Restrict Impact	If online programs stop people from doing daily activities, they have to restrict the time of using the internet.	Restrict
3-Parents and schools should permit children to use the internet only for learning.	Permit Restrict	Parents and schools should permit children to use the internet freely.	Permit
4- Without rules for using the internet, it cannot be beneficial for children.	Beneficial Precious	Having rules for using the internet can make it beneficial to learning.	Beneficial
5-The internet is interesting for children because it provides diverse types of online games.	Diverse Previous Innovative Beneficial	Some schools have special programs for children with diverse learning problems.	Diverse
6-Schools can use the internet as an innovative system for learning.	Innovative Precious Beneficial	Online learning and group work are not innovative ideas anymore because these days, many schools use them.	Innovative
7-The internet is a precious tool for children's development if they use it in a suitable way.	Precious Innovative beneficial	Children should be taught that time is precious and it should not be spent on computer games.	Precious

All acceptable answers were listed in Appendix 18 as the answer keys for Section 3 of Lesson 1 and 2. In Appendix 18, the answer key of Lesson 2 shows that there are some unrevised sentences within which more than one target word can be matched (i.e., sentences 8 and 13). These sentences were not changed because, after revising other sentences and preventing interchangeable use of words, other options were ruled out. For instance, two target words,

namely *enforce* and *restrict*, could make sense in sentence 8. However, just one of them, *enforce*, was acceptable because this word could not be used in sentence 7 instead of *restrict*. The same was true for sentence 13 in which more than one word might fit the context, but only one word was acceptable due to the impossibility of interchangeable use.

Based on the third issue about gap-filling tasks explained by Read (2012), decisions had to be made about the importance of grammar and spelling. Because in this study, the gap-filling tasks were used as part of the treatment to induce evaluation, not as a test, and the participants could check their performance at the end of the task by looking at the lessons in which they encountered the words, grammar and spelling were not taken into account. In addition, because the focus of the treatment was knowledge of the meanings of the target words, grammar and spelling were not considered.

Apart from considering the three issues with gap-filling tasks, two other faults were noticed in Section 3. First, in this section, after evaluating each target word, the participants had to express their ideas about the sentences by circling Yes, No or Not sure (Appendix 1, pp. 158-159 and pp. 165-166). As observed in the pilot study, the participants were so deeply involved in evaluating the target words that they did not give attention to expressing ideas. For the main study, the learners were not asked to express their ideas and Yes, No, or Not sure options were deleted from the task. In addition, in the instruction, the participants had to be asked to use each word only once. The final instruction as presented in Appendix 17 (pp. 196-197 and pp. 203-204) is as follows:

‘Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about changes in weather/using the internet that may be different from your idea(s) about it. You can use each word only once’.

For the productive retrieval groups, this instruction also informed the participants that the first letter of each word was provided. For the receptive retrieval groups, the instruction stated that the words were provided at the bottom of the page.

The last change to this section was eliminating the extra writing activity. As mentioned before, it had been predicted that the composition-writing group could finish Section 3 faster than the productive retrieval group. The extra time of the latter group could be filled by assigning a short paragraph writing task after the gap-filling activity on a topic different from the topic of the lesson. However, after the pilot study, it was found out that the time taken to

do Section 3 was not very different for the two groups. Doing Section 3 of Lesson 1 took 20 minutes for the composition-writing group while the productive retrieval group completed the same section in 17 minutes. Section 3 of Lesson 2 took 22 minutes for the composition-writing group and 18 minutes for the productive retrieval group. The reason that productive retrieval group took a long time to complete Section 3 was that the participants needed time to comprehend each sentence and remember a word they had encountered in the lessons with the given first letter that could match the context. Thus, for the main study, this extra writing activity was eliminated for the retrieval groups.

3.3.2.3 Piloting the Post-tests

In the pilot study, for each test, the papers were collected when the participants had completed the test. The time given to each test was recorded as shown in Table 3.13.

Table 3.13

Time of tests for each group

	The first group	The second group
Lesson 1:		
Productive test	10 mins	9 mins
Translation test	7 mins	8 mins
Matching test	10 mins	12 mins
Lesson 2:		
Productive test	12 mins	10 mins
Translation test	9 mins	9 mins
Matching test	11 mins	12 mins

3.3.2.3.1 Productive test

In the productive test, a list of L2 definitions had been provided and the participants were instructed to write an L2 word for each definition. The first letter of each word was provided to avoid eliciting words other than the target and non-essential words with similar meanings. After piloting the productive test of Lesson 1, it was found out that the words *destruction* and *devastating* that start with D and have similar definitions can be confusing for the participants despite the fact that they are not the same part of speech. In the productive test, some of the participants wrote one of the words instead of the other. Since these words were linked to similar concepts, it was decided to replace one of them with another word. The word *destruction* in Section 2 of Lesson 1 was replaced with the word *contamination* (Appendix 17,

p. 195). Since the paragraph in which *contamination* appears is about the effect of human activities and natural factors on weather change, this word fits the paragraph well. The new word was chosen according to the criteria for choosing other target and non-essential words. It is an abstract noun within the 4000 word frequency list. The definition provided for this word in the tests is 'making something dirty'.

Another point was noticed by one of the raters while scoring the productive tests. The target words *fundamental* in Lesson 2 and *vital* in Lesson 1 shared a similar meaning. For the groups who received the meanings of words, the definitions provided for *fundamental* and *vital* were 'basic' and 'necessary' respectively. Thus, these groups might not have been aware that the words shared a similar meaning. However, the groups who looked up the meanings of words might find out that, in addition to 'basic', *fundamental* meant 'essential' and, in this sense, had a meaning similar to the meaning of *vital*. This could lead to different conditions for learning the two target words across the groups. To provide a similar learning condition, one of the target words could be eliminated from the study.

As mentioned above, in the the pre-test, one noun, one verb, and one adjective, namely *global, temperature,* and *alert,* were eliminated from Lesson 1 due to being familiar to most of the participants. From Lesson 2, only one verb, *decrease,* was deleted. By deleting *fundamental* from Lesson 2, it was possible to avoid having two synonymous words and to make the lessons more balanced in terms of the number and class of the target words.

In addition to identifying problems in the test, the productive test papers were marked by two raters. The aim was to pilot the procedures for scoring the tests. In marking the tests, spelling was not taken into account, but the participants were expected to show their knowledge of the words. The marks were also used to calculate the reliability of the test (explained in 3.3.2.4.2).

3.3.2.3.2 Translation test

In the translation test of each lesson, a list of the target and non-essential words had been given to the participants and they were asked to provide their L1 meanings. The participants were instructed to provide only one and the best L1 meaning for each L2 word. Part of speech was not taken into account in judging the answers.

The translation tests were marked by two raters that were different from those who marked the pre-test. They had not seen the answer key and had not talked about the acceptable answers in advance. Inter-rater reliability was calculated based on the marks they gave to the

participants independently (as explained in 3.3.2.4.2). Then, in order to make a list of acceptable answers to be used in the main study, the raters used an answer key that had been compiled after marking the vocabulary pre-test (Appendix 16). The participants' responses to the translation test items were judged by two raters. On some occasions, the raters had to reach an agreement on including a response in the answer key. Then, a list of acceptable answers was prepared as the final answer key and was added to Appendix 16.

3.3.2.3.3 Matching test

The matching tests were piloted considering four issues about piloting matching tests mentioned by Read (2012, p. 309). First, there should be only one correct response for each item. Second, the correct answers should be expressed clearly and not able to be guessed easily by being more specific compared to other options. The third issue is that wrong options should function effectively by attracting some of the test-takers' attention. The final issue about piloting a matching test is that the items should be an appropriate level of difficulty. These issues had provided the basis for writing the matching test specifications and designing the test accordingly. Note taking during the pilot study and analysing the participants' papers helped to check the function of each test item and option.

Regarding the first issue, the pilot study confirmed that each item on the matching test had only one correct response. Considering the second issue, when designing the test for this study, definitions had been kept short and clear. To avoid giving grammatical clues, the three or four words presented together had shared the same part of speech because the participants might identify the correct answers based on their knowledge of the word class. The wrong definitions had defined words with the same part of speech as the target and non-essential words. The wrong definitions had had some words in common with correct definitions and mostly had had similar lengths. Definitions had been arranged by length and there had been no pattern for the responses. The pilot study revealed that none of the options was noticeably more specific than others. However, there was an option that had not been expressed clearly. As mentioned by some of the participants, option C shown in Figure 3.3 was not clear enough. In their view, it could not match any of the words. Option C, 'to handle a problem', had been supposed to define the word *confront*, but it did not function properly. This problem had remained unnoticed in piloting the productive test. Surprisingly, a few participants matched this option with the correct word. However, to make it clear for all participants, option C had

to be revised. In all instructional and test materials, the definition of *confront* was changed to 'to deal with a problem or difficult situation'.

18-Rescue B 19- Persuade H 20- Confront C 21- Endeavour E	A. to make things easier B. to save somebody form danger C. to handle a problem D. to put something out of sight E. to try very hard to do something F. to think carefully about something G. to find detailed information about something H. to make somebody do something by giving good reasons

Figure 3.3

An unclear option in the sixth cluster of the matching test of Lesson 1

The third issue about piloting the matching tests was to identify the options that had not been chosen by any of the participants. Table 3.14 presents such options in Lesson 1.

Table 3.14

Unattractive options of Lesson 1

Options	
A	• Fishing • to make things easier
B	• Health • happening in a specific place
C	• very heavy
E	• very delicious

Not choosing some of the options could be due to the large number of them in each set. It seemed that six and eight options for three and four items respectively were too many for the participants. In addition, some of the options that were long and needed a lot of reading caused frustration among the participants. Thus, in addition to replacing each of the options mentioned in Table 3.14, some others were deleted. In replacing the options, the test specifications and other options in their sets were considered. The replacements were as follows:

When designing the test, the first word, 'fishing', had been provided as an option because it resembled the second option, 'farming', which was the correct response as shown in

Figure 3.4. The participants ignored this option probably due to its irrelevance to the topic and the content of the lesson. This option was changed to ‘warning’, which was relevant to the content of the lesson and had some letters in common with option B.

1- Climate G	A. fishing
2-Phenomena H	B. farming
3-Agriculture B	C. strange stories
4-Contamination F	D. safe condition
	E. making something new
	F. making something dirty
	G. general weather condition
	H. events that are not fully understood

Figure 3.4

Revising the first cluster of the matching test of Lesson 1

The next problematic case is option A, ‘to make things easier’, shown in Figure 3.5. This option was replaced by ‘to stop internet overuse’.

18-Rescue C	A. to make things easier
19- Persuade H	B. to put something out of sight
20- Confront G	C. to save somebody form danger
21- Endeavour D	D. to try very hard to do something
	E. to think carefully about something
	F. to find detailed information about something
	G. to deal with a problem or difficult situation
	H. to make somebody do something by giving good reasons

Figure 3.5

An unattractive option in the sixth cluster of the matching test of Lesson 1

Two of the B options shown in Figures 3.6 and 3.7 had been unsuccessful in attracting the participants’ attention.

5- Temperature F	A. growth
6- Emission D	B. health
7- Expansion A	C. the production of energy
	D. the production and sending out of gas
	E. a unit that shows how wet or dry a place is
	F. a measure of how hot or cold a place or thing is

Figure 3.6

An unattractive option in the second cluster of the matching test of Lesson 1

8-Gradual A	A. happening slowly over a long time
9-Global E	B. happening in a specific place
10-Devastating D	C. causing a lot of questions
11-Blazing H	D. causing a lot of harm
	E. for the whole world
	F. for local people
	G. very angry
	H. very hot

Figure 3.7

An unattractive option in the third cluster of the matching test of Lesson 1

'Health' was changed to 'earth' and 'happening in a specific place' to 'happening after a long time'. Both new options had some commonalities with other neighbouring options and were related to the content of the lesson.

The final problems in the matching test of Lesson 1 were depicted in Figure 3.8. Option E was changed to 'naturally' and option C was replaced with 'general'. The word 'very' was eliminated from options D and F to make them shorter.

12- Numerous B	A. Most
13- Exceptional D	B. Many
14- Vital F	C. very heavy
	D. very unusual
	E. very delicious
	F. very necessary

Figure 3.8

Revising the fourth cluster of the matching test of Lesson 1

The matching test of Lesson 2 had just one unattractive option. No participant had chosen option C shown in Figure 3.9. This option was changed into 'not having a change in life', which was relevant to the content of the lesson.

5- Deficiency D	A. having a problem
6-Addiction F	B. having a close relationship
7-Intimacy B	C. not having access to technology
	D. not having enough of something
	E. being able to do necessary things
	F. being unable to stop doing something

Figure 3.9

An unattractive option in the second cluster of the matching test of Lesson 2

After making all the changes and deletions, the options were arranged by length. Appendix 21 presents the matching tests of Lessons 1 and 2 that were used in the main study.

To control the level of difficulty of test items, which is the last issue for piloting a matching test, all the target and non-essential words had been chosen from the 3000 and 4000 word frequency levels. All the definitions had included words within the 1000 and 2000 word frequency levels, which were assumed to be known by intermediate learners.

3.3.2.4 Validity and reliability of the tests

3.3.2.4.1 Validity

The validity of the tests was demonstrated by following the steps suggested by Schmitt (2010 b, p. 181). According to him, validity must be shown through test development and performance. In this study, at the development stage, for each test, detailed specifications had been developed based on the literature and the tests content had been specified through the following statements:

1- The tests measured a specific set of words taught through two lessons.
2- Test items were abstract nouns, verbs, and adjectives at the 3000-4000 frequency levels.
3- The test items measured three aspects of knowledge of words presented through two lessons. Productive test items measured active recall, translation test items measured passive recall, and matching test items measured passive recognition of words.
4- There were also specifications for each of the tests, as explained above.

At the performance stage (the pilot study), notes were taken on the test administration procedures, the participants' performance, and their understanding of what they were supposed

to do in order to identify problems and to make modifications. Teacher and rater comments were also taken into account in finalizing the teaching and testing materials.

3.3.2.4.2 Reliability

The inter-rater reliability of each productive test taken by the groups is shown in Table 3.15. In calculating the inter-rater reliability, the marks given by two raters before their meeting to discuss the answers were taken into account.

Table 3.15

Inter-rater reliability of the productive tests

Productive tests	Cronbach's Alpha	Number of raters
Lesson 1 taken by the first group	.983	2
Lesson 2 taken by the first group	.992	2
Lesson 1 taken by the second group	.995	2
Lesson 2 taken by the second group	.972	2

The values of Cronbach's Alpha for all the tests were very high, indicating that the raters were highly consistent with each other in their marking.

The inter-rater reliability of each translation test taken by the groups is presented Table 3.16. The same as the productive tests, the marks given by raters before their conference were taken into account. The values of Cronbach's Alpha show the reliability of scoring.

Table 3.16

Inter-rater reliability of the translation tests

Translation tests	Cronbach's Alpha	Number of raters
Lesson 1 taken by the first group	.996	2
Lesson 2 taken by the first group	.997	2
Lesson 1 taken by the second group	.983	2
Lesson 2 taken by the second group	.985	2

The matching test reliability was also measured by calculating Cronbach's Alpha. To do this, the same as in Schmitt, Schmitt, and Clapham's (2001) study, each cluster of the matching tests was considered as an item. Each of the matching tests of Lessons 1 and 2 included 6 clusters and each cluster had either 3 or 4 items (Appendix 5). The clusters were marked based on the number of items answered correctly. The clusters in which no item was answered correctly

received zero. The clusters with four items received .25 for each correct answer. The clusters with three items received .33 for each correct response. The clusters in which all the items were answered correctly received 1.

After marking the clusters, the matching tests of Lessons 1 and 2 were combined and Cronbach's Alpha was calculated for all 12 clusters and reported in Table 3.17. The reliability coefficient was above the accepted minimum of .70 (Riazi, 2016, pp. 271 and 74).

Table 3.17

Reliability of the matching test

	Cronbach's Alpha	Confidence Interval	Number of items (Clusters)	Number of Participants
All Groups	.764	95%	12	23

3.4 Main Study

3.4.1 Participants

After the pilot study and revision of the instructional and test materials, the main study was carried out. For the main study, 10 groups of participants had to be chosen (as presented in Table 3.1). As in the pilot study, the participants' grades on the X-Lex test had to be double-checked with their grades in their previous semester test. Because assessing the groups' level of proficiency based on their previous semester grades was only possible if they had taken the same or at least comparable tests, two branches of the same language institute were chosen as the sites for the main study. In both branches, the same as the pilot study, the second edition of the American English File (AEF) was the teaching material and learners had to take the AEF tests at the end of each semester. Thus, all the procedures for choosing the participants (the X-Lex and AEF tests) were the same as in the pilot study.

The teachers who worked for the institute had classes in both branches. The Consent Forms (CFs) and the Participants' Information Sheets (PISs) were distributed among teachers of intermediate classes and in total, three teachers, two from one branch and one from the other, gave consent to participate. In both branches, teachers of intermediate classes had to teach part of the book 3A of the AEF series. The teachers were consulted about the suitability of the research material for their classes and they stated that the material was appropriate for their

students' level of proficiency. Each of the teachers distributed the CFs and the PISs among the learners of their own classes and all the learners gave consent to participate in the study. Overall, ten classes of the two branches of the language institute with learners above the age of 18 participated in the main study. The number of learners in each class was between 16 and 19.

Most of the participants had been studying English in the same language institute for several semesters. They had been placed in the intermediate class based on the scores they had achieved in the AEF test in the previous semester. Four participants had lately joined the language institute and been placed in the intermediate classes. In both branches, for placing new learners in an appropriate class, first they had to have an interview with the institute's supervisors as a pre-assessment. Then, based on their interview, the learners were recommended to take the AEF test of a specific level as a placement test. The four new learners had taken an AEF test that was the same as the one the other participants had passed to be placed in the intermediate classes. In order to identify the outliers among the participants as a whole, the X-Lex test was administered to all ten classes and the scores were reviewed. In addition to taking the X-Lex test, the participants had to answer a question about their language background. As in the pilot study, some of the participants stated that they knew Azari and written Arabic. Since these languages could not interfere with English vocabulary learning, no one was excluded from the study due to their language background.

The participants' AEF test scores in the previous semester were obtained from the language institute to be double-checked with their X-Lex test scores. The participants who had been studying English in the same institute had report cards showing their previous performance in the AEF test. For the four new students, the results of the placement tests (the AEF test) were used for double-checking.

To identify the outliers, first, the students' X-Lex test scores were assessed. A larger number of the participants' scores were a little below 2500-3250, extending down to what Milton (2009) defines as the A2 level. It seemed unreasonable to exclude all these potential participants, so a lower boundary was set. The scores equal to or above 2400 and equal to or below 2650 were considered as insiders because this range (2400-2650) included the most frequent scores and the score frequency dropped significantly below 2400 and above 2650. The participants whose X-Lex scores were not within this range were judged on the basis of their AEF test scores. If their AEF test scores were within three standard deviations of their group's mean, they were not be considered as outliers; otherwise, their data were eliminated from the study. In calculating the groups' means, the participants whose X-Lex scores were not within

the boundaries (outliers) were eliminated in order to prevent their scores from having a disproportionate effect on the groups' means. The descriptive statistics for the groups are presented in Table 3.18.

After the calculations, the data of 6 participants out of 174 were excluded from the study (two in Group 5, one in each of Groups 7 and 8, and two in Group 9). One of the excluded participants in Group 5 had a high pre-test score showing that she knew more than 10 of the 37 target and non-essential words. Although she was not identified as an outlier based on her X-Lex and the AEF test scores, her data were eliminated from the study due to her unexpected knowledge of the words compared to that of other participants.

Table 3.18

Descriptive statistics of the AEF test scores of all groups

	No.	Minimum	Maximum	Mean	Std. Deviation
Group 1 (no search and receptive retrieval)	17	88.00	93.00	90.23	1.82
Group 2 (no search and productive retrieval)	17	86.00	94.00	90.82	2
Group 3 (search and receptive retrieval)	17	87.00	93.00	89.70	1.86
Group 4 (search and productive retrieval)	16	88.00	93.00	91	1.50
Group 5 (no search and sentence writing)	17	86.00	93.00	90.17	2.15
Group 6 (no search and composition writing)	17	88.00	93.00	90.35	1.36
Group 7 (search and sentence writing)	18	87.00	94.00	90.50	1.97
Group 8 (search and composition writing)	17	86.00	94.00	89.88	2.49
Group 9 (search and no evaluation)	16	89.00	93.00	90.75	1.29
Group 10 (control group)	16	86.00	94.00	89.81	2.07

After identifying the outliers and eliminating their data, the means of the groups' X-Lex test scores and the AEF test scores were calculated and shown in table 3.19. Then, the means of the groups' means in the X-Lex test and in their AEF test were calculated and were shown in tables 3.20 and 3.21 to compare the groups in terms of their level of proficiency.

Table 3.19

The groups' means in the X-Lex and AEF tests

	X-Lex test means	*AEF test* means (Out of 100)
Group 1	2511.76	90.23
Group 2	2523.52	90.82
Group 3	2497.05	89.70
Group 4	2531. 25	91.00
Group 5	2547.05	90.17
Group 6	2505.88	90.35
Group 7	2502.77	90.50
Group 8	2541.17	89.88
Group 9	2528.12	90.75
Group 10	2525.00	89.81

As Tables 3.20 and 3.21 show, all the groups' means in the X-Lex and AEF tests are within two standard deviations above and below the mean of the groups' means, which indicates the groups' homogeneity.

Table 3.20

Descriptive statistics for the combined groups in the X-Lex test

No.	Minimum	Maximum	Mean	Std. deviation
10	2497.05	2547.05	2521.35	16.60

Table 3.21

Descriptive statistics for the combined groups in the AEF test

No.	Minimum	Maximum	Mean	Std. deviation
10	89.70	91.00	90.32	.447

3.4.2 Pre-test

The vocabulary pre-test (Appendix 2) was administered to measure the participants' initial knowledge of the target and non-essential words. The average time spent on the test for all ten groups was 12.8 minutes. Each participant's pre-test paper was scored by two raters to record the participant's knowledge of the target and non-essential words so that in the post-tests, the words that had already been known by each of the participants could be subtracted from his/her

test scores. To make such records and to maintain consistency in judging the participants' knowledge, the raters reviewed the participants' papers and remarked them using the answer key designed based on the pilot study (Appendix 16). In specific cases (35 items in the pre-tests taken by 168 participants), they had to reach an agreement about whether or not to accept the answers. At the final stage, all the target and non-essential words known by each participant were recorded for later use.

Table 3.22 presents an overview of the groups' knowledge of 25 target and 12 non-essential words after the final judgement.

Table 3.22

The groups' known target and non-essential words

	Known target words			Known non-essential words		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum
Group 1	0	.588	4*	0	.235	1
Group 2	0	.058	1	0	.588	3*
Group 3	0	.147	3	0	.647	2
Group 4	0	.687	2	0	.875	3*
Group 5	0	1.67	3	0	.294	1
Group 6	0	.647	2	0	.352	2
Group 7	0	.666	3	0	.333	2
Group 8	0	.470	3	0	.882	3*
Group 9	0	.875	3	0	.625	2
Group 10	0	.750	4*	0	.625	1

As Table 3.22 shows, in each of the Groups 1 and 10, where the maximum knowledge of the target words was observed, only one participant knew 4 target words (shown by an asterisk). In each of the Groups 2, 4, and 8, where the maximum knowledge of the non-essential words is shown, only one participant knew 3 non-essential words (shown by asterisks). With the exception of these particular participants, the groups did not know more than 3 of the 12 non-essential words and more than 4 of 25 target words.

3.4.3 Treatment and Procedures

In the first session of data collection, the nine experimental groups and the control group took the homogeneity test and the vocabulary pre-test. In the second session, the experimental groups received the first lesson. The third session was for administering three vocabulary post-

tests of Lesson 1 to all ten groups. In the fourth session, the experimental groups received Lesson 2 and in the fifth session, all the groups took three vocabulary post-tests of Lesson 2. The above-mentioned procedures are summarized in Table 3.23. For all the experimental groups, the time between each lesson and the relevant tests were kept the same (two days).

Table 3.23

Data collection procedures for the main study

	Session 1	Session 2	Session 3	Session 4	Session 5
Experimental groups	X-Lex and vocabulary pre-test	Lesson 1	Vocabulary post-tests of Lesson 1	Lesson 2	Vocabulary post-tests of Lesson 2
Control group	X-Lex and vocabulary pre-test		Vocabulary post-tests of Lesson 1		Vocabulary post-tests of Lesson 2

Carrying out the procedures and administering the tests were done by the teacher of each class. They were trained to use the instructional materials and to administer the tests. The researcher acted as an observer to monitor the proper use of the materials and tests. As presented in Table 3.1, the nine experimental groups experienced different combinations of the involvement factors with different degrees. The control group, who did not receive the lessons, only took the X-Lex test, vocabulary pre-test, and post-tests.

Time on task for different parts of the lessons varied across groups. For each part/section, time on task was the interval between starting the task and completing it by the participants. The difference between the maximum and minimum time of completion across the groups can be explained by the fact that the groups who had to look up the meanings of words did the tasks more slowly than the groups who were given the meanings. Type of evaluation could also affect time on task for different groups. Table 3.24 presents time on task for parts of Section 1.

Table 3.24

Time on task for Section 1 of the lessons

	Groups who had to look up the words' meanings (Groups 3, 4, 7, 8, and 9)	Groups who were given the words' meanings (Groups 1, 2, 5, and 6)
Lesson 1	Average time on task (times are approximate)	Average time on task (times are approximate)
Section 1:		
Part A	4.5 mins	
Part B	12.5 mins	9.5 mins
Part C	12.5 mins	8.5 mins
Part D	13 mins	11.5 mins
Total time	42.5 mins	34 mins
Lesson 2		
Section 1:		
Part A	6.5 mins	
Part B	15 mins	11.5 mins
Part C	13 mins	10.5 mins
Part D	11.5 mins	9.5 mins
Total time	46 mins	38 mins

Time on task for completing Section 2 was between 6 and 11 minutes depending on the participants' motivation to look up the meanings of the non-essential words. In each of the lessons, this section included a table that had to be filled out based on the information provided in a paragraph.

Time on task for completing Section 3 of the lessons was more varied across the groups, given that each group had to complete one of the four task types. This section in Lessons 1 and 2 included 12 and 13 words respectively.

As shown in Table 3.25, the groups who had to write a composition spend more time on Section 3 than the other groups who had to write sentences or to fill in the gaps of the sentences (receptive and productive retrieval).

Table 3.25

Time on task for Section 3

	Average time on task (times are approximate)
Receptive retrieval groups	14 mins
Productive retrieval groups	17.5 mins
Sentence writing groups	20 mins
Composition writing groups	21.5 mins

The normal class time in the language institute was an hour and a half. As given in Table 3.24, Section 1 of the lessons took at most 46 minutes. Section 2 and 3 took at most 11 and 21.5 minutes respectively. Thus, the maximum time given to the lessons was 78.5 minutes, which meant that the normal class time was sufficient for carrying out all of the experimental treatments.

3.4.4 Post-tests

The ten groups of participants took three post-tests, namely productive, translation, and matching tests. The approximate times for taking the tests across the groups are shown in Table 3.26.

Table 3.26
Time of tests

	Average time on test (times are approximate)
Productive test	10.5 mins
Translation test	8 mins
Matching test	12 mins

3.4.4.1 Productive and Translation Post-tests

3.4.4.1.1 Scoring

The productive and translation post-tests were scored subjectively through the following steps to determine each participant's score:

1- The papers were marked by two raters using the answer key. One point was given to the answers the raters considered correct, half a point to partially correct answers, and zero to wrong or no answers.
2- The discrepancies between the raters were identified and the participants' answers to the test items were judged using the answer key designed based on the pilot study. In some cases where the responses had not been listed in the answer key, the raters had to reach an agreement about the acceptability of the answers.
3- Each participant was given a score.
4- The words that each participant had already known (based on the pre-test) were subtracted from his/her score. The resultant score showed the target and non-essential

words each participant learned. There were a few cases where the answers were accepted by the raters in the pre-test, but the participants provided incorrect or no answers in the translation test. These cases were not counted in the participants' scores.

5- For each participant, the scores of the target and non-essential words were separated.
6- For each participant, for the target words, the average score in the productive tests of lessons 1 and 2 and in the translation tests of Lessons 1 and 2 was calculated.
7- The non-essential words learned by each participant, if any, were recorded to be analysed separately.

3.4.4.2 Matching Tests

3.4.4.2.1 Scoring

The matching test was scored objectively. Compared to the productive and transltion tests, fewer steps were taken to determine each participant's score:

1- The papers were marked by the researcher. One point was given to the correct answers and zero to the wrong or no answers.
2- For each participant, known words were subtracted from the test score.
3- The scores for the target and non-essential words were separated.
4- For each participant, for the target words, the average score in the matching test of lessons 1 and 2 was calculated.
5- The non-essential words learned by each participant, if any, were recorded to be analysed separately.

The recorded scores were entered in the SPSS files for statistical analyses. As explained in Chapter 4, descriptive statistics, including mean scores, standard deviations, normality and homogeneity of variance, were calculated. The results of normality and homogeneity of variance tests contributed to choosing suitable inferential statistics for comparing the groups.

Chapter 4: Results

4.1 Descriptive Statistics

After administering the tests, the scores were analysed and presented in the form of descriptive statistics. Table 4.1 presents descriptive statistics of the tests for all groups. As the table shows, the mean score of the control group (Group 10) is the lowest and the mean scores of the groups who experienced sentence-writing and composition-writing (Groups 5, 6, 7, and 8) are the highest. The descriptive statistics provided the basis for further analyses, which are presented in the following sections.

Table 4.1

Descriptive statistics of the post-tests

	Productive Tests (Active Recall)		Translation Tests (Passive Recall)		Matching Tests (Passive Recognition)	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Group 1	1.29	.70	2.11	1.22	3.23	1.59
Group 2	2.48	.93	3.11	.94	4.05	.84
Group 3	2.10	1.15	3.73	.96	4.29	1.10
Group 4	2.68	.573	3.81	1.04	4.40	.89
Group 5	3.22	1.003	4.30	1.67	6.05	2.18
Group 6	3.16	.814	5.04	1.83	6.14	1.80
Group 7	3.20	1.179	4.54	2.30	5.22	1.28
Group 8	3.94	1.79	4.89	1.49	5.52	1.41
Group 9	.593	.48	1.92	.98	3.15	.88
Group 10	.25	.43	.34	.38	.21	.51

4.2 Normality and Homogeneity of Variance

To identify the appropriate inferential statistics for comparing the groups, checks were made to see whether the score data met the assumptions of parametric tests, including interval scale, independence, normal distribution, and homogeneity of variances (Larson-Hall, 2010, pp. 74-75; Riazi, 2016, p. 229). The assumptions of interval scale and independence had already been

met. Normality and homogeneity of variance were examined and presented in the following paragraphs.

Normality of data was assessed by inferential statistics (Shapiro-Wilk test), skewness and kurtosis, and visual means (histograms). The Shapiro-Wilk test (Appendix 22) indicated that some of the datasets were not normally distributed ($p \leq .05$).

Skewness and kurtosis were divided by their standard errors and the obtained values between -2 and +2 show normal distribution with 95 percent confidence (Bachman, 2004, p. 74). A more conservative approach is to consider any values of skewness and kurtosis within -1 and +1 as an indication of normal distribution (Phakiti, 2015, p. 38). The values were summarized and shown in Appendix 22.

Histograms were evaluated in terms of their conformity to a normal distribution (Appendix 23). Table 4.2 summarizes the findings of each approach of assessing normality. The datasets that, based on all three approaches, conformed to a normal distribution are shown in the right column.

Table 4.2

Normality of the Datasets

	Not normal				**Approximately normal**
	Shapiro-Wilk test	skewness and kurtosis	skewness and kurtosis/Std. errors	Histogram	
Productive test (Active Recall)	8, 9, 10	8, 9, 10	10	3, 4, 6, 9, 10	1, 2, 5, 7
Translation test (Passive Recall)	1, 7, 10	1, 4, 5, 6	1	1, 2, 4, 6, 7, 9, 10	3, 8
Matching test (Passive Recognition)	2, 4, 6, 9, 10	4, 6, 10	10	1, 3, 4, 5, 6, 9, 10	7, 8

Homogeneity of variance of the groups that had to be compared was detected by looking at side-by-side boxplots (Appendix 24) and by running the Levene's test (Appendix 25). In assessing the boxplots, the boxes with similar lengths indicate that there is no substantial difference between variances (Larson-Hall, 2010, p. 87). In choosing the statistics resulting from the Levene's test, the median instead of the mean was used, which according to Larson-Hall (2016, p. 122) results in a more robust test.

The homogeneity of variance of some of the datasets confirmed by the Levene's test was ruled out by visual examination of the boxplots. An example was the homogeneity of variance of the productive test scores for Groups 1 and 6. The Levene's test results also ruled out the homogeneity of variance of some of the datasets confirmed by visual examination of the boxplots. An example was the homogeneity of variance of the matching test scores for Groups 2 and 5. Only datasets whose homogeneity of variance was confirmed by both approaches were taken into account. They are shown in Table 4.3.

Table 4.3

Homogeneity of variance of the datasets

	Productive test	Translation test	Matching test
Datasets with homogeneity of variance	1 and 5		2 and 3
	2 and 6		3 and 5
	2 and 9		5 and 7

After assessing the normality and homogeneity of variance of the datasets, it was revealed that only a small number of datasets met all the criteria for applying a parametric test (groups 1 and 5's productive test scores). Thus, it was decided to apply a non-parametric test even in the case of these data sets that met the criteria.

4.3 Inferential Statistics

In order to answer the first research questions, which investigates the contribution of each type of evaluation to vocabulary learning, six comparisons were made, including Groups 1-2, 5-6, 2-5, 2-6, 1-5, and 1-6. To answer the second research question, which investigates the contribution of search to vocabulary learning, alone and in combination with types of evaluation, another six comparisons were made, including Groups 1-9, 2-9, 1-3, 2-4, 5-7 and 6-8. To provide additional evidence regarding the contribution of search, it was necessary to make complementary comparisons, including Groups 2-3, 3-5, and 3-6. The control group (Group 10) was also compared with one of the experimental groups (Group 9) that had the lowest mean in all post-tests to show the effects of the treatment on the groups. It is worth adding that the focus of all the stated comparisons was the target words for which the need factor had a fixed degree (moderate). To answer the third research question, which investigates

the contribution of each degree of the need factor to vocabulary learning, the non-essential words were analysed and discussed at the end of this chapter.

The Mann-Whitney U test was applied to compare the groups. Because multiple comparisons were made among the groups, Bonferroni correction was applied to each set of comparisons in order to obtain a corrected *p* value for the set. A corrected *p* value is calculated by dividing the *p* value (.05) by the number of performed statistical analyses (Larson-Hall, 2010, p. 252; Maxwell & Delaney, 2004, p. 202). The observed *p* values lower than the corrected *p* values are considered statistically significant and shown by asterisks in the following tables. In all the tables, II stands for the Involvement Index, U for the Mann-Whitney U test, Z for the z score, and R for the Effect size. The involvement factors were abbreviated: N stands for need, S for search, and E for evaluation.

4.3.1 The Contribution of Each Type of Evaluation to Vocabulary Learning

To investigate the contribution of each type of evaluation to vocabulary learning, the following groups were compared: 1-2, 5-6, 2-5, 2-6, 1-5, and 1-6. Because six multiple comparisons were made, the corrected *p* value is .0083 (.05/6).

As shown in Table 4.4, Groups 1 and 2 experienced the same degrees of need (+) and search (-). The only difference between the groups was the type of evaluation induced by the treatment. Based on the predictions of the Involvement Load Hypothesis (ILH), Group 2's treatment should be more effective that Group 1's treatment, given that the involvement index (II) of the former is higher. Comparing the groups may show any difference between the receptive and productive retrieval in terms of their impact on the vocabulary learning in the absence of search.

Table 4.4

Productive, Translation, and Matching Test statistics for comparing Groups 1 and 2

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	11.76	47	-3.388	.581	.001*	.0083
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	23.24					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	12.62	61.5	-2.876	.493	.004*	.0083
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	22.38					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	13.97	84.5	-2.101	.360	.036	.0083
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	21.03					

The Mann-Whitney U tests indicated that the difference between the groups in the productive and translation tests was significant. In other words, in the absence of search, productive retrieval with a significantly higher mean rank was more effective for vocabulary learning (as shown in the productive and translation tests) than receptive retrieval, which conforms to the ILH predictions. However, when it comes to the matching test, although the productive retrieval group had a higher mean rank, the difference between the groups was not statistically significant, which is in contrast with the ILH predictions. The effect sizes are medium ($.30 \leq r \leq .49$) to large ($r \geq .50$).

Groups 5 and 6 experienced the same degrees of need (+), search (-), and evaluation (++). The only difference between the groups is the type of evaluation. While Group 5 experienced sentence-writing, Group 6 experienced composition-writing. Both treatments have the same II and, as the ILH predicts, they should be equally effective for vocabulary learning. The groups were compared to investigate any difference between sentence-writing and composition-writing in terms of their impact on vocabulary learning in the absence of search.

Table 4.5

Productive, Translation, and Matching test statistics for comparing Groups 5 and 6

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	17.68	141.5	-.104	.017	.917	.0083
Group 6	N +	S -	E (composition-writing) ++	II=3	17	17.32					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	15.47	110	-1.190	.204	.234	.0083
Group 6	N +	S -	E (composition-writing) ++	II=3	17	19.53					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	17.26	140.5	-.139	.023	.890	.0083
Group 6	N +	S -	E (composition-writing) ++	II=3	17	17.74					

The Mann-Whitney U tests indicated that the difference between the groups in all three tests was not significant. In other words, in the absence of search, sentence-writing and composition-writing were equal in terms of their impact on vocabulary learning, which conforms to the ILH predictions. Although the groups were not significantly different, the mean rank of the group who experienced composition-writing was higher than that of the group who experienced sentence-writing in the translation test. The effect sizes are all small ($R \leq .30$).

Groups 2 and 5 who experienced the same degree of need (+), search (-), and evaluation (++) were compared. Although the groups have the same II, they experienced different types of evaluation. While Group 5 experienced sentence-writing, Group 2 experienced productive retrieval. As the ILH predicts, the treatment should be equally effective for vocabulary learning. The groups were compared to investigate any difference between sentence-writing and productive retrieval in terms of their impact on vocabulary learning in the absence of search.

Table 4.6

Productive, Translation, and Matching test statistics for comparing Groups 2 and 5

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	13.94	84	-2.101	.360	.036	.0083
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	21.06					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	13.29	73	-2.472	.424	.013	.0083
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	21.71					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	12.29	56	-3.086	.529	.002*	.0083
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	22.71					

The Mann-Whitney U tests indicated that the difference between the groups was only significant in the matching test. In other words, in the absence of search, sentence-writing showed no significantly higher mean rank compared to productive retrieval in the productive and translation tests, which conforms to the ILH predictions. However, even in these tests, the mean rank of sentence-writing was higher. The effect sizes are medium ($.30 \leq r \leq .49$) to large ($r \geq .50$).

Groups 2 and 6, who experienced the same degree of need (+), search (-), and evaluation (++), were compared. The groups have the same II, but while Group 6 experienced composition-writing, Group 2 experienced productive retrieval. As the ILH predicts, the treatments should be equally effective for vocabulary learning. The groups were compared to investigate any difference between productive retrieval and composition-writing in the absence of search.

Table 4.7

Productive, Translation, and Matching test statistics for comparing Groups 2 and 6

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	13.97	84.5	-2.091	.358	.037	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	21.03					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	12.32	56.5	-3.040	.521	.002*	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	22.68					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N+	S-	E (productive retrieval) ++	II=3	17	11.71	46	-3.444	.590	.001*	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	23.29					

The Mann-Whitney U tests showed that the difference between the groups was significant in the translation and matching tests, but not the productive test. In the absence of search, composition-writing with a higher mean rank was more effective than productive retrieval in terms of its impact on vocabulary learning (as shown in the translation and matching tests), which does not conform the ILH predictions. Even in the productive test, the mean rank of composition-writing was higher. The effect sizes are medium ($.30 \leq r \leq .49$) to large ($r \geq .50$).

Groups 1 and 5 experienced the same degrees of need (+) and search (-). However, the groups' treatments induced different types and degrees of evaluation. Groups 1 and 5 experienced receptive retrieval and sentence-writing respectively. Based on the ILH assumptions, Group 5's treatment with a higher II should be more effective for vocabulary learning. The groups were compared to investigate any difference between sentence-writing and receptive retrieval in terms of their impact on vocabulary learning in the absence of search.

Table 4.8

Productive, Translation, and Matching test statistics for comparing Groups 1 and 5

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	9.76	13.00	-4.546	.779	.000*	.0083
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	25.24					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	11.41	41.00	-3.581	.614	.000*	.0083
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	23.59					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	11.41	41.00	-3.582	.614	.000*	.0083
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	23.59					

The Mann-Whitney U tests indicated that the differences between the groups in all three tests were significant. In other words, in the absence of search, sentence-writing was more effective for vocabulary learning than receptive retrieval, which conforms to the ILH predictions. The effect sizes are large ($r \geq .50$).

Groups 1 and 6 experienced the same degrees of need (+) and search (-). However, while Group 1 experienced receptive retrieval, Group 6 experienced composition-writing. The ILH predicts that Group 6's treatment with higher II should be more effective for vocabulary learning. The groups were compared to investigate any difference between composition-writing and receptive retrieval in terms of their impact on vocabulary learning in the absence of search.

Table 4.9

Productive, Translation, and Matching test statistics for comparing Groups 1 and 6

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N+	S-	E (receptive retrieval) +	II=2	17	9.59	10.00	-4.658	.798	.000*	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	25.41					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N+	S-	E (receptive retrieval) +	II=2	17	10.47	25.00	-4.131	.708	.000*	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	24.53					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N+	S-	E (receptive retrieval) +	II=2	17	10.88	32.00	-3.892	.667	.000*	.0083
Group 6	N+	S-	E (composition-writing) ++	II=3	17	24.12					

The Mann-Whitney U tests showed that the difference between the groups in all three tests was significant. In other words, in the absence of search, composition-writing with a significantly higher mean rank was more effective for vocabulary learning than receptive retrieval, which conforms to the ILH predictions. The effect sizes are all large ($r \geq .50$).

In summary, investigating the differences between the types of evaluation revealed that first, on the one hand, productive retrieval was more effective than receptive retrieval, as shown by the active and passive recall tests. Even in the matching test (passive recognition test), where the difference between the groups was not significant, productive retrieval showed a higher mean rank. This may indicate that productive retrieval contributed more than receptive retrieval to vocabulary development, especially word recall. On the other hand, both sentence-writing and composition-writing were significantly more effective than receptive retrieval. Thus, receptive retrieval was the least powerful type of evaluation for vocabulary learning.

Second, on the one hand, comparing productive retrieval and sentence-writing indicated that they were significantly different only for passive recognition of words. However, in all tests, the mean ranks of sentence-writing were higher. On the other hand, composition-writing was significantly more effective than productive retrieval as shown in the passive recall and recognition tests. Even in the productive test, the mean rank of composition-writing was higher. Thus, in terms of the effectiveness for vocabulary learning, productive retrieval had the middle status between receptive retrieval and writing.

Third, direct comparison of sentence-writing and composition-writing indicated that they were not significantly different, although composition-writing had a higher mean rank in the translation test. However, by comparing the above two paragraphs, which compared productive retrieval with each of sentence-writing and composition-writing, it can be concluded that composition-writing may be more effective than sentence-writing for vocabulary learning. The reason is that the difference between composition-writing and productive retrieval (in the passive recall and recognition tests) is greater than the difference between sentence-writing and productive retrieval (in the passive recognition test).

4.3.2 The Contribution of Search to Vocabulary Learning

The contribution of search to vocabulary learning was investigated in two steps: first, the contribution of search alone and second, search in combination with types of evaluation.

4.3.2.1 The Contribution of Search Alone

This step includes two comparisons. Group 9, who experienced search but did not experience evaluation at all, was compared with Groups 1 and 2, who experienced receptive and productive retrieval without search. Groups 1 and 2 had the lowest means among the groups who experienced any kind of evaluation and no search. Thus, any detected difference between Group 9 and Groups 1 and 2 could be generalized to other experimental groups experiencing evaluation and no search. Because two comparisons were made, the corrected *p* value is .025 (.05/2).

Groups 1 and 9 experienced the same degree of need (+). However, the groups' treatments induced different degrees of search and evaluation. Group 1 experienced receptive retrieval without search, whereas Group 9 had search and no evaluation. Based on the ILH assumptions, the groups' treatments should be equally effective for vocabulary learning. The

groups were compared to investigate the contribution of search and evaluation in the form of receptive retrieval to vocabulary learning.

Table 4.10

Productive, Translation, and Matching test statistics for comparing Groups 1 and 9

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	21.56	58.50	-2.832	.493	.005*	.025
Group 9	N +	S +	E -	II=2	16	12.16					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	16.85	133.50	-.091	.015	.927	.025
Group 9	N +	S +	E -	II=2	16	17.16					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	16.68	130.50	-.201	.034	.841	.025
Group 9	N +	S +	E -	II=2	16	17.34					

Table 4.10 showed that the difference between the groups in the translation and matching tests was not significant. In other words, the contribution of search and evaluation in the form of receptive retrieval could be similar for passive recall and recognition of words, which conforms to the ILH predictions. However, when it came to the productive test, the difference between the groups was significant ($p \leq .025$), and receptive retrieval showed a higher mean rank, which contradicts the ILH predictions. In other words, for productive vocabulary learning the contribution of evaluation in the form of receptive retrieval was higher than search. The effect sizes are small ($R \leq .30$) to medium ($.30 \leq r \leq .49$).

Groups 2 and 9 experienced the same degree of need (+). However, the groups' treatments induced different degrees of search and evaluation. While Group 2 experienced

productive retrieval without search, Group 9 had search and no evaluation. Based on the ILH assumptions, Group 2's treatment should be more effective for vocabulary learning. The groups were compared to investigate the contribution of search and evaluation in the form of productive retrieval to vocabulary learning.

Table 4.11

Productive, Translation, and Matching test statistics for comparing Groups 2 and 9

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	24.68	5.50	-4.734	.824	.000*	.025
Group 9	N +	S +	E -	II=2	16	8.84					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	21.91	52.50	-3.024	.526	.002*	.025
Group 9	N +	S +	E -	II=2	16	11.78					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	21.79	54.50	-3.023	.526	.002*	.025
Group 9	N +	S +	E -	II=2	16	11.91					

The Mann-Whitney U tests showed that the difference between the groups in all three tests was significant. In other words, the contribution of evaluation in the form of productive retrieval was considerably higher than search for vocabulary learning. The effect sizes are all large ($r \geq .50$).

To sum up, in comparing search and evaluation in the form of receptive retrieval, it was revealed that they could have similar impacts on passive recall and recognition of words (as shown in the translation and matching tests). However, when it came to the productive test, where the difference was significant, evaluation appeared superior over search for all types of

vocabulary learning measured in this study (active recall, passive recall, and passive recognition). Since sentence and composition-writing groups had higher means compared to the productive retrieval groups' means, this result can be generalized to other groups and it can be concluded that evaluation in the form of sentence and composition-writing is superior to search for all types of vocabulary learning measured in this study.

4.3.2.2 The Contribution of Search in Combination with Types of Evaluation

In the following paragraphs, the combination of search with each type of evaluation was compared with evaluation alone. Four comparisons were made, including 1-3, 2-4, 5-7 and 6-8. Therefore, the corrected *p* value is .0125 (.05/4)

Groups 1 and 3 both experienced receptive retrieval (+) and moderate need (+). However, the degree of search for the groups was different. While for the former, search was absent, for the latter, the treatment involved search (Table 4.12). Based on the ILH assumptions, Group 3's treatment with higher II should be more effective for vocabulary learning. The groups were compared to investigate how the presence of search can influence the contribution of receptive retrieval to vocabulary learning.

Table 4.12

Productive, Translation, and Matching test statistics for comparing Groups 1 and 3

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	14	85	-2.072	.355	.038	.0125
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	21					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	11.44	41.50	-3.572	.612	.000*	.0125
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	23.56					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 1	N +	S -	E (receptive retrieval) +	II=2	17	13.71	80	-2.239	.384	.025	.0125
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	21.29					

As the table presented, the difference between Groups 1 and 3 was significant in only the translation test, which is in accordance with the ILH predictions. The effect sizes are medium ($.30 \leq r \leq .49$) to large ($r \geq .50$). The analyses indicated that search could increase the effectiveness of receptive retrieval for only passive recall of words.

Groups 2 and 4, who both experienced productive retrieval (++) and moderate need (+), were different in terms of the degree of search induced by the treatments (Table 4.13). Since the II of Group 4's treatment is higher than that of Group 2, as the ILH predicts, it should be more effective for vocabulary learning. The groups were compared to investigate how the presence of search affects the contribution of productive retrieval to vocabulary learning.

Table 4.13

Productive, Translation, and Matching test statistics for comparing Groups 2 and 4

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	15.76	115	-.766	.133	.444	.0125
Group 4	N +	S +	E (productive retrieval) ++	II=4	16	18.31					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	14.24	89	-1.703	.296	.089	.0125
Group 4	N +	S +	E (productive retrieval) ++	II=4	16	19.94					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	15.15	104.5	-1.176	.204	.240	.0125
Group 4	N +	S +	E (productive retrieval) ++	II=4	16	18.97					

As presented in Table 4.13, no significant difference was found between Groups 2 and 4's treatments. Contrary to the ILH predictions, productive retrieval combined with search showed no significantly higher mean rank compared to productive retrieval without search. The effect

sizes are small (r ≤ .30). The analyses showed that search could not contribute to the effectiveness of productive retrieval for vocabulary learning.

Groups 5 and 7 both experienced sentence-writing (++) and moderate need (+). However, the degree of search was different across the groups (Table 4.14). Based on the ILH assumptions, Group 7's treatment with higher II should be more effective for vocabulary learning. The groups were compared to investigate how the presence of search can influence the contribution of sentence-writing to vocabulary learning.

Table 4.14

Productive, Translation, and Matching test statistics for comparing Groups 5 and 7

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	17.62	146.5	-.216	.036	.829	.0125
Group 7	N +	S +	E (sentence-writing) ++	II=4	18	18.36					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	18.47	145	-.265	.044	.791	.0125
Group 7	N +	S +	E (sentence-writing) ++	II=4	18	17.56					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 5	N +	S -	E (sentence-writing) ++	II=3	17	20.29	114	-1.298	.219	.194	.0125
Group 7	N +	S +	E (sentence-writing) ++	II=4	18	15.83					

As the Mann-Whitney U tests showed, the difference between Groups 5 and 7 in all three tests was not significant. Contrary to the ILH predictions, sentence-writing combined with search did not show significantly higher mean ranks compared to sentence-writing without search. The effect sizes are all small (r ≤ .30). The analyses indicated that search could not affect the contribution of sentence-writing to vocabulary learning.

Groups 6 and 8 who both experienced composition-writing (++) and moderate need (+) were different in terms of the degree of search induced by the treatment (Table 4.15). Since the II of Group 8's treatment is higher than that of Group 6, as the ILH predicts, it should be more effective for vocabulary learning. The groups were compared to investigate how the presence of search affects the contribution of composition-writing to vocabulary learning.

Table 4.15

Productive, Translation, and Matching test statistics for comparing Groups 6 and 8

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 6	N +	S -	E (composition-writing) ++	II=3	17	15.65	113	-1.092	.187	.275	.0125
Group 8	N +	S +	E (composition-writing) ++	II=4	17	19.35					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 6	N +	S -	E (composition-writing) ++	II=3	17	18.15	133.5	-.380	.065	.704	.0125
Group 8	N +	S +	E (composition-writing) ++	II=4	17	16.85					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 6	N +	S -	E (composition-writing) ++	II=3	17	19.18	116	-.988	.169	.323	.0125
Group 8	N +	S +	E (composition-writing) ++	II=4	17	15.82					

The Mann-Whitney U tests showed that the difference between Groups 6 and 8 in all three tests was not significant. Contrary to the ILH predictions, composition-writing combined with search did not show significantly higher mean ranks compared to composition-writing without search. The effect sizes are all small ($r \leq .30$). The analyses indicated that search did not affect the contribution of composition-writing to vocabulary learning.

In sum, the analyses indicated that search could only increase the effectiveness of receptive retrieval for passive recall of words and had no impact on the effectiveness of productive retrieval, sentence-writing, and composition-writing.

4.3.2.3 Additional Comparisons to Investigate the Contribution of Search to Receptive Retrieval

For detailed investigation of the contribution of search to receptive retrieval, additional comparisons were made. The combination of search and receptive retrieval was compared with each of productive retrieval, sentence-writing, and composition-writing. The comparisons include Groups 2-3, 3-5, and 3-6. The corrected *p* value is .0166 (.05/3)

Groups 2 and 3 experienced the same degrees of need (+). However, the degrees of search and type of evaluation were different across the groups. While Group 2's treatment induced productive retrieval (++) and no search (-), Group 3's treatment involved receptive retrieval (+) and search (+). Although the involvement factors and their degrees are different across the groups, both treatments have the same II and, as the ILH predicts, they should be equally effective for vocabulary learning. The groups were compared to estimate the effect of search in increasing the power of receptive retrieval for vocabulary learning compared to productive retrieval (Table 4.16).

Table 4.16

Productive, Translation, and Matching test statistics for comparing Groups 2 and 3

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	19.12	117	-.957	.164	.339	.0166
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	15.88					
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	14.68	96.5	-1.666	.285	.096	.0166
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	20.32					
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 2	N +	S -	E (productive retrieval) ++	II=3	17	16.68	130.5	-.493	.084	.622	.0166
Group 3	N +	S +	E (receptive retrieval) +	II=3	17	18.32					

As the results of the Mann-Whitney U tests showed, in accordance with the ILH predictions, there was no significant difference between Groups 2 and 3's treatments. In other words, productive retrieval was as effective as the combination of receptive retrieval and search for vocabulary learning. The effect sizes are all small ($r \leq .30$).

Groups 3 and 5 experienced the same degrees of need (+). However, search and evaluation were varied across the groups. While Group 3 experienced search and receptive retrieval, Group 5's treatment involved sentence-writing without search. Both treatments have the same II and, as the ILH predicts, they should be equally effective for vocabulary learning. The groups were compared to estimate the effect of search in increasing the power of receptive retrieval for vocabulary learning compared to sentence-writing (Table 4.17).

Table 4.17

Productive, Translation, and Matching test statistics for comparing Groups 3 and 5

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	13.09	69.50	-2.594	.444	.009*	.0166
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	21.91					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	15.03	102.5	-1.456	.249	.145	.0166
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	19.97					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	12.94	67.00	-2.690	.461	.007*	.0166
Group 5	N+	S-	E (sentence-writing) ++	II=3	17	22.06					

The Mann-Whitney U tests indicated that the difference between the groups was significant in the productive and matching tests. In other words, sentence-writing was significantly more effective than the combination of receptive retrieval and search for active recall and passive recognition, which does not conform to the ILH predictions. The effect sizes are small ($R \leq .30$) to medium ($.30 \leq r \leq .49$).

Groups 3 and 6 experienced the same degrees of need (+). However, Group 3 experienced search and receptive retrieval and Group 6 experienced composition-writing without search. Both treatments have the same II and, as the ILH predicts, they should be equally effective for vocabulary learning. The groups were compared to estimate the effect of search in increasing the power of receptive retrieval for vocabulary learning compared to composition- writing (Table 4.18).

Table 4.18

Productive, Translation, and Matching test statistics for comparing Groups 3 and 6

Productive test **(Active recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	13.03	68.50	-2.641	.453	.008*	.0166
Group 6	N+	S-	E (composition-writing) ++	II=3	17	21.97					
Translation test **(Passive recall)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	13.85	82.50	-2.145	.367	.032	.0166
Group 6	N+	S-	E (composition-writing) ++	II=3	17	21.15					
Matching test **(Passive recognition)**					Descriptive Statistics		Mann-Whitney U				
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)	Corrected *p*
Group 3	N+	S+	E (receptive retrieval) +	II=3	17	12.56	60.50	-2.910	.499	.004*	.0166
Group 6	N+	S-	E (composition-writing) ++	II=3	17	22.44					

Table 4.18 showed that the difference between the groups was significant in the productive and matching test, not in the translation test. In other words, composition-writing was more effective than the combination of receptive retrieval and search for active recall and passive recognition, which conforms to the ILH predictions. The effect sizes are all medium ($.30 \leq r \leq .49$).

In sum, the analyses indicated that search could increase the power of receptive retrieval to the same level as the power of productive retrieval for all types of vocabulary learning measured in this study. However, search could not increase the power of receptive retrieval to the same level as the power of sentence and composition-writing for active recall and passive recognition of words.

4.3.3 The Difference between the Control and Experimental Groups

In addition to the preceding analyses, Group 9, which had the lowest mean in all the tests (as highlighted in Table 4.1) was compared with the control group (Group 10) as shown in Table 4.19. If this group performed significantly better than the control group, then it can be assumed that the other groups had better results as well. Since only one comparison was made, *p* value is .05.

Table 4.19

Productive, Translation, and Matching test statistics for comparing Groups 9 and 10

Productive test (Active recall)					Descriptive Statistics		Mann-Whitney U			
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)
Group 9	N +	S +	E -	II=2	16	19.75	76	-2.081	.367	.037*
Group 10	N -, +, or ++	S -, +, or ++	E -, +, or ++	II=0-5	16	13.25				
Translation test (Passive recall)					Descriptive Statistics		Mann-Whitney U			
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)
Group 9	N +	S +	E -	II=2	16	23.22	20.5	-4.103	.725	.000*
Group 10	N -, +, or ++	S -, +, or ++	E -, +, or ++	II=0-5	16	9.78				
Matching test (Passive recognition)					Descriptive Statistics		Mann-Whitney U			
					N	Mean Rank	U value	Z	*R*	Sig. (2-tailed)
Group 9	N +	S +	E -	II=2	16	24.44	1	-4.938	.873	.000*
Group 10	N -, +, or ++	S -, +, or ++	E -, +, or ++	II=0-5	16	8.56				

As presented in Table 4.19, the differences between the groups in all three tests are significant. The effect sizes are medium ($.30 \leq r \leq .49$) to large ($.50 \leq r \leq .1.0$). The results indicate that all the experimental groups are significantly different from the control group, and that shows that all the treatments had some positive effect on vocabulary learning.

4.3.4 The Contribution of Need (Analysing Learning the Non-essential Word)

Each participant's knowledge of the non-essential words had already been measured by the pre-test and recorded. Subsequently, the participants encountered the non-essential words in

Section 2 of each lesson, which included a table that had to be filled in based on the information provided in a paragraph. However, it was not necessary to know the meanings of the non-essential words embedded in the paragraph for filling in the table. The participants had been allowed to use a monolingual dictionary while doing the task and had been asked to circle any word they had looked up in a dictionary. After each lesson, the participants' papers were scrutinized to record the non-essential words each participant circled. Not circling the non-essential words was taken as evidence that the participants had not noticed them or had not needed them.

The non-essential words ignored by the participants at the time of task completion (not circled), did not receive any involvement index, given that the need and search factors were absent. Circling the words was taken as evidence that the participant decided to pay attention to the words so that the need factor would be strong (++) and the search factor would appear (+).

There is uncertainty about the degree of evaluation. If the participants only looked up the meanings of words in the classroom, the evaluation factor could not be induced. However, if the participants remembered the words after the lessons and practiced them in speaking and writing or encountered them in listening and reading materials outside the research context, evaluation could appear to a strong or a moderate degree. Thus, for these words, the involvement factor and their degrees could be between 3 and 5 depending on the degree of evaluation, as shown in Table 4.20. The following paragraphs explain how the research results have been interpreted considering the uncertainty about the degree of evaluation.

Table 4.20

The non-Essential words' involvement indices

	Need	Search	Evaluation	Involvement Index
Non-essential words	- or ++	- or +	-, +, or ++	3-5

Careful study of the participants' papers showed that presentation of non-essential words in Lesson 1 did not alert the participants to look for them in Lesson 2 because not all of the participants who circled the non-essential words in Lesson 1 circled them in Lesson 2 as well. Many of the participants who ignored them in Lesson 1 ignored them in Lesson 2 as well. Many of them circled words other than the non-essential words in both lessons. If presenting the non-essential words in Lesson 1 had alerted the participants to look for them in Lesson 2, the degree

of need for the non-essential words of Lesson 2 would have been different. Since in this condition, need was imposed by the task, it would have been moderate (+), not strong (++).

In the post-tests administered after each lesson, the non-essential words were scored separately. Only the non-essential words the participants had circled while doing the tasks were taken into account in the post-tests in order to eliminate any variable that could affect the degree of need. For each participant, the non-essential words that were already known (based on the pre-test) were excluded from the test scores.

Some non-essential words circled by the participants in the lessons produced wrong or no answers in the tests. Altogether, 85 participants in nine experimental groups showed learning of about 189 non-essential words (100 in Lesson 1 and 89 in Lesson 2) in the post-tests. Some of the participants answered the words correctly in all post-tests. Some of the words were answered correctly in only one or two of the tests. In total, 167 words in the matching tests, 87 words in the translation tests, and 23 words in the productive tests were answered correctly, as presented in Table 4.21.

Table 4.21

The non-essential words circled and/or answered correctly in the post-tests

	The number of times the words had been circled	Correctly answered in the Matching test	Correctly answered in the Translation test	Correctly answered in the Productive test
Numerous	36	25	12	4
Aim	34	19	13	3
Blazing	32	19	14	6
Intimacy	29	17	6	1
Irritating	28	12	3	0
Participate	27	13	7	1
Contamination	26	20	13	4
Rescue	23	11	6	3
Acquire	19	8	3	0
Expansion	18	11	3	1
Illogical	18	7	1	0
Endeavour	14	5	4	0
Total number	**304**	**167**	**87**	**23**

The 85 participants who received scores (even zero) for learning the non-essential words formed a group (the non-essential word group) that could be compared with the other groups who learned the target words (target words groups). Three issues had to be considered before

making comparisons. First, the number of participants in the non-essential word group was substantially higher than the number of participants in any individual target word group. To deal with this issue, the participants of the non-essential word group were randomly divided into five. The reason for having five groups was that the majority of the target word groups included 17 participants. By dividing the non-essential word group into five groups (Groups A, B, C, D, and E), each group could include 17 participants, which is a similar size to the rest of the groups in the study. The descriptive statistics of the non-essential word groups are presented in Table 4.22. The second issue was the number of target and non-essential words. The target word groups encountered 12 target words in Lesson 1 and 13 in Lesson 2. For each participant, the average score in Lessons 1 and 2 was calculated. Therefore, the maximum score in any test of the target words was 12.5. The number of the non-essential words presented in Section 2 of each lesson was 6. For each participant, the scores of non-essential words of Lessons 1 and 2 were added. Thus, the maximum score in any test of the non-essential words for each participant could be 12, which was comparable to the target word scores. The last issue is related to the comparison of the groups and the assumptions of the statistical tests. The participants in the non-essential word groups also belonged to one of the experimental groups (target word groups). Therefore, in comparing each of the Groups A, B, C, D, and E with each of the experimental groups, some of the participants could be in both groups, but not all of the participants in the groups were the same. Since by making these comparisons the assumptions of none of the statistical tests could be met, the comparisons were made merely based on descriptive statistics.

Table 4.22

The non-essential word groups' (A, B, C, D, and E) descriptive statistics

	Productive Tests (Active recall)		Translation Tests (Passive recall)		Matching Tests (Passive recognition)	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Group A	.294	.587	1.235	1.714	1.882	1.996
Group B	.388	.607	1.222	1.352	2.222	1.733
Group C	.294	.685	.882	.992	1.705	2.284
Group D	.058	.242	.411	.618	1.941	1.853
Group E	.117	.332	1.058	1.434	1.647	1.966

As Table 4.22 shows, the highest mean in the productive test belongs to Group B (.388). The highest means in the translation and matching tests belong to Groups A and B (1.235 and 2.222 respectively). These groups were compared with the experimental (target word) groups and control group in the study. If any difference is found, then, it can be said that the other non-essential word groups are also different from the experimental and control groups. Tables 4.23, 4.24, and 4.25 present the comparison of the groups' means in the productive, translation, and matching tests respectively.

Table 4.23

Comparing the means of Group B and other groups in the productive test

Group B's mean	The experimental and control groups' means									
	1	2	3	4	5	6	7	8	9	10
.388	1.29	2.48	2.10	2.68	3.22	3.16	3.20	3.94	.593	.25

Table 4.24

Comparing the means of Group A and other groups in the translation test

Group A's mean	The experimental and control groups' means									
	1	2	3	4	5	6	7	8	9	10
1.235	2.11	3.11	3.73	3.81	4.30	5.04	4.54	4.89	1.92	.34

Table 4.25

Comparing the means of Group B and other groups in the matching test

Group B's mean	The experimental and control groups' means									
	1	2	3	4	5	6	7	8	9	10
2.222	3.23	4.05	4.29	4.40	6.05	6.14	5.22	5.52	3.15	.21

As Tables 4.23, 4.24, and 4.25 show, the means of the selected non-essential word groups (A and B) in all tests are lower than those of all the experimental groups. Supposing that the degree

of evaluation was zero for the non-essential word groups, the results suggest that learning the non-essential words involving search and strong need (with II= 3) was substantially inferior to learning the target words involving search and moderate need (e.g., experimental Group 9 with II= 2). In other words, in the presence of search and absence of evaluation, the degree of need may not matter for vocabulary learning and need with a moderate degree is probably sufficient for learning.

Supposing that the degree of evaluation was moderate for the non-essential word groups, then, the results show that learning the non-essential words involving search, strong need, and moderate evaluation (II= 4) was substantially inferior to learning the target words involving search, moderate need, and moderate evaluation (e.g., experimental Group 3 with II= 3). Even if the degree of evaluation was strong for non-essential word groups, then, the results indicate that learning the non-essential words involving search, strong need, and strong evaluation (II= 5) was substantially inferior to learning the target words involving search, moderate need, and strong evaluation (e.g., experimental Groups 4, 7, and 8 with II= 4). In other words, in the presence of search and any degree and type of evaluation, the degree of need (moderate or strong) may not matter for vocabulary learning.

The results also suggest that learning the non-essential words involving strong need, search, and any degree of evaluation was only superior to the control group who did not experience the treatment.

In sum, investigating the difference between the non-essential word groups and the experimental and control groups shows that strong need and search combined with any degree and type of evaluation are less effective than any other combinations of moderate need, search, and evaluation (contrary to the ILH predictions). Strong need may not be as effective as search and evaluation for vocabulary learning.

Chapter 5: Discussion

5.1 The Involvement Load Factors in This study

The Involvement Load Hypothesis (ILH) proposed by Laufer and Hulstijn (2001) introduced three factors, namely need, search, and evaluation as the effective factors in vocabulary-focused tasks. Need was given three degrees: minus, when a word is not needed for task completion; moderate, when learners are asked to learn a word or when learning a word is necessary for task completion; and strong, when learners themselves decide to learn a word. Search was given two degrees: minus, when the meanings of words are provided for learners; and plus, when learners are required to look up the meanings of words in a dictionary. The last factor, evaluation, was given three degrees: minus, when learners do not have to make a comparison which word or sense of a word to choose or use; moderate, when learners have to make such a comparison (e.g., in gap-filling tasks); and strong, when learners need to use a word in an original context (e.g., writing sentences or paragraphs). Later, Nation and Webb (2011, p. 4) added the concept of retrieval to search so that search could have three degrees: minus, when the meanings of words are given; moderate, when learners are required to look up the meanings of words or to retrieve the meanings of words (receptive retrieval); and strong, when learners are required to look up or retrieve the forms of words (productive retrieval). In this study, the two concepts of look up and retrieval were separated, as mentioned in Chapter 1 to be able to account for the differences between memory retrieval and dictionary search in terms of their impact on vocabulary learning and to be able to extend the evaluation factor as mentioned in 1.5.

In the following sections, where each research question is discussed in turn, E refers to evaluation that can be of four types: receptive memory retrieval (R), productive memory retrieval (P), sentence-writing (SW), and composition-writing (CW) respectively. S refers to dictionary search.

5.2 Research Questions

5.2.1 Research Question 1

1- What is the contribution of each type of evaluation to vocabulary learning?

1a: Based on the answer to the first research question, is it possible to give more degrees to the evaluation factor?

To answer the first research question, the types of evaluation were compared in the absence of search and the results are discussed as follows:

First, it was found out that the contribution of SW, CW, and P (varieties of strong evaluation) are higher than R (moderate evaluation) to vocabulary learning except that for passive recognition of words, no significant difference was detected between P and R. Overall, the superiority of SW, CW, and P (with high involvement indices) over R (with a lower index) is in accordance with the ILH's assumption that tasks with higher involvement indices are more effective for vocabulary learning. The results can also be explained by the Retrieval Effort Hypothesis (Pyc & Rawson, 2009), which states if retrievals are done successfully, difficult retrievals result in superior learning compared to easier ones. According to Ellis and Beaton (1993 b) and Nation (2013, p. 51), retrieving information from memory in order to write a sentence or a composition or to provide the forms of words (productive retrieval) is more difficult than retrieving their meanings (receptive retrieval). Thus, based on the Retrieval Effort Hypothesis, the former group of retrievals should result in superior learning, which is supported by the findings in this study.

The superiority of P over R for active recall and their similarity for passive recognition of words can be explained by Nation's (2013, pp. 56, 458) and Mondria and Wiersma's (2004, p. 85) view that productive learning best suits productive use and that productive learning leads to the developing of some receptive knowledge. The first part supports the finding that the group who experienced productive memory retrieval could outperform the group who experienced receptive memory retrieval in the active recall test, which involved supplying the target words' forms. The second part explains the reason that the productive retrieval group was similar to the receptive retrieval group in the passive recognition test, which involved recognizing the target words' forms. The similarity of P and R for passive recognition of words may also show that for this aspect of vocabulary learning, which according to the hierarchy of difficulty proposed by Laufer, Elder, Hill, and Congdon (2004) and Laufer and Goldstein (2004) is the easiest type of learning, it made no difference whether the learners were involved in P (strong evaluation) or R (moderate evaluation). When more difficult learning was involved (active and passive recall of words), the superiority of P over R appeared, which is in accordance with the claims put forward by the Retrieval Effort Hypothesis.

As mentioned before, CW and SW had the same involvement index as P and all of them involved demanding retrievals. Because no similarity was found between CW and R and between SW and R, it may be argued that there had to be no similarity between P and R as well. This can be accounted for by a characteristic of CW and SW, which makes them different

from P. The use of words in original contexts in CW and SW, which is absent in P, may make them more powerful for vocabulary learning and different from R for all the kinds of vocabulary learning measured in this study. This is in line with Joe's (1998) findings that using words in original contexts enhances vocabulary learning and, the higher the level of production, the greater the opportunity for vocabulary learning. Nation (1993, p. 121) also stated that generative use of new words through constructing original sentences leads to superior vocabulary learning compared to using new words in contexts provided by someone else. It may show that P, SW, and CW should not be given the same involvement index. The next paragraphs, which discuss the comparisons between P, SW, and CW, provide supports for giving different indices to them.

Second, compared with P, SW is more effective for passive recognition and CW for passive recall and recognition. This goes against the ILH assumption that activities with the same involvement indices should be equally effective for vocabulary learning. The superiority of SW and CW over P was in the lower and middle points of the hierarchy of difficulty (Laufer et al., 2004; Laufer & Goldstein, 2004), showing that for the most difficult learning (active recall), these varieties of strong evaluation were similar. This may indicate that the production involved in all these types of evaluation could similarly contribute to active recall of words, which is in accordance with Nation's (2013, p. 458) and Mondria and Wiersma's (2004, p. 85) view that productive learning (or form recall) best suits productive use. Nevertheless, when it comes to passive recall and recognition of words, it seems that different types of strong evaluation act differently. SW and CW could contribute more than P to passive recall and recognition of words. This indicates that CW, SW, and P should not be given the same involvement index (++).

Third, the comparison of CW and SW revealed no significant difference between them in terms of their impact on vocabulary learning. The ILH and the Retrieval Effort Hypothesis claims underpin this finding, given that CW and SW have the same involvement index and both involve difficult memory retrieval. However, previous studies provided contradictory evidence regarding the difference in the effectiveness of SW and CW for vocabulary learning. For instance, while Zou (2017) indicated that CW is significantly more effective than SW, Kim (2011) provided evidence showing that these two output activities are equally effective. Zou's (2017) argument for her findings is that, because CW requires creating an original context for each target word as well as associating the contexts of the target words coherently, it induces a higher involvement load compared to SW, which only requires creating a separate original context for each target word. Nonetheless, Kim (2011) explains her results by referring to a

statement by Laufer (personal communication, March 15, 2005) that the difficulty (the higher level of cognitive processing) of CW, which results from maintaining coherence, does not necessarily show a higher level of processing of new words. In both SW and CW, learners process new words in original contexts, which leads to similar vocabulary learning. The results of the present study could provide additional evidence in favour of Kim's and Laufer's claims for the similarity of SW and CW for vocabulary learning. It is worth adding that comparing P with each of SW and CW indicated that the difference between CW and P (passive recall and recognition) was greater than the difference between SW and P (passive recognition). It may show that CW could be more effective than SW. However, because in the direct comparison of CW and SW no significant difference was found, it can be concluded that the difference between SW and CW, if any, is not substantial enough to make a difference in their power for vocabulary learning and thus to give them two different involvement indices.

Putting all the comparisons together, if R involves a moderate degree of evaluation (+), P, which is stronger than R, may involve strong evaluation (++). Accordingly, SW and CW, which are the strongest ones, may involve very strong evaluation (+++). This conclusion may modify the degrees of the evaluation factor of the ILH. Instead of having three degrees of prominence for E (-, +, and ++), it is possible to have four degrees (-, +, ++, +++). Having the new degrees for the evaluation factor may affect the predictions of the ILH mentioned previously. Based on the new degrees given to the evaluation factor, tasks that used to have the same involvement indices, and based on the ILH predictions had to be equally effective for vocabulary learning, do not share the same involvement index anymore and their different effects on vocabulary learning can be explained using new degrees. In other words, differentiating between CW and SW on the one hand and P on the other hand will contribute to extending the evaluation factor and having more accurate estimate of task effectiveness.

In addition, in comparing P with each of R, SW, and CW, it was revealed that the task involvement index does not always show the effectiveness of task for all types of vocabulary learning. The reason is that P, with a higher involvement index, compared to R, had a similar impact on passive recognition of words. At the same time, P, with a lower index (if the suggested degrees are applied) compared to SW and CW, had a similar impact on active recall of words. The conclusion might be that, in judging task effectiveness for vocabulary learning, beside involvement indices, type of evaluation and type of vocabulary learning should be taken into consideration. The characteristics of each type of evaluation, such as using words in original contexts (in SW and CW), production of form without using words in original contexts

(in P), and evaluation without production of form (in R) were shown to have their own specific impact on each type of vocabulary learning.

To summarize the discussion regarding the first research question and Question 1a the following statements were made:

1- CW and SW, which involve using words in original contexts, are the most effective types of evaluation for vocabulary learning and can be given the equal involvement index of three (very strong degree of evaluation).
2- On the continuum of the effectiveness for vocabulary learning, P has the middle position between CW and SW on the one end and R on the the other. It can be given the involvement index of two (strong degree of evaluation).
3- R is the least effective type of evaluation and can be given the involvement index of one (moderate degree of evaluation).
4- Although P is different from R in terms of its degree of prominence, its effectiveness for the easiest type of learning (passive recognition of words) is similar to R. Only in learning more difficult aspects of vocabulary (active and passive recall) does the superiority of P over R appear.
5- Although CW and SW are different from P in terms of their degrees of prominence (if the suggested degrees are given to them), the production involved in these three types of E could be similarly effective for active recall of words. However, SW and CW contribute more than P to learning other aspects of vocabulary (passive recall and recognition).

5.2.2 Research Question 2

2- What is the contribution of search to vocabulary learning?

2a: Based on the answers to the first and second research questions, it is possible to detect the impact of the distribution of the factors on vocabulary learning?

To investigate the contribution of S to vocabulary learning, two series of comparisons were made. The first group focused on the contribution of S alone to vocabulary learning by comparing it with different types of E. The second group investigated the contribution of S to vocabulary learning when it is combined with different types of E. The latter group of comparisons could also reveal how S might affect the contribution of different types of E to vocabulary learning.

5.2.2.1 The Contribution of Search Alone

The results of the comparisons between different types of evaluation and search are discussed as follows:

First, SW, CW, and P (varieties of strong evaluation) were more effective than S to vocabulary learning. This is in accordance with the ILH assumptions (based on both the original and new degrees), given that the involvement indices of SW, CW, and P are higher than S.

Second, with respect to the difference between R (moderate evaluation) and S, the results revealed that for active recall, the contribution of R was significantly greater than the contribution of S. However, going down the hierarchy of difficulty toward passive recall and recognition, R and S became similar in terms of their contribution to vocabulary learning. In other words, for the most difficult kind of learning (active recall), a moderate degree of evaluation contributed more than dictionary search. For easier learning, such as passive recall and recognition, it made no difference which of R or S was incorporated.

The conclusion drawn from comparing R and S can be put together with the conclusion from comparing P and R in the first research question. It seems that by going up the hierarchy of difficulty (from passive recognition to active recall of words), evaluation and the production involved in it gained importance. The reason is that the superiority of P, which involves production of form, over R, which involves no production of form, appeared in more difficult types of vocabulary learning (active and passive recall). Likewise, the superiority of R, which does not involve production of form, but some degree of evaluation, over S, which involves neither production nor evaluation, appeared in learning the most difficult aspect of vocabulary (active recall).

The above-mentioned results may provide a more detailed explanation for Keating's (2008) statement about the S factor of the ILH. As he mentioned, S may involve some degree of E, given that looking up the meanings of words in a dictionary involves choosing from several senses of a word, which induces moderate E and helps learners to focus on form. His view is supported by cognitive psychology, which relates retention of information to the depth of processing (e.g., Craik & Lockhart, 1972). Laufer and Hill (2000), referring to the depth of processing, state that retention of words hinges substantially on deep processing and elaboration (attention to the features of words and their relations to other words); thus, effective vocabulary tasks are those that result in elaboration. In Laufer and Hill's (2000) view, using dictionaries to look up words is one of such effective tasks. The results of the current study,

while not contradicting Keating's and Laufer and Hill's view about the function of S, indicated that even if S involves elaboration, the degree of elaboration was less powerful than the degree of elaboration involved in E of any type (SW, CW, P and R). As the results indicated, all varieties of strong E were more effective than S. Even R, which involves moderate evaluation, was more effective than S for active recall of words. In other words, Keating's view that S involves moderate evaluation was not supported by the result of this study. Although, as Keating stated, S may help learners to focus on form, this cannot compete with the focus on form achieved through R (as measured by active recall test, which involved retrieving forms of words).

These conclusions are in contrast with Barcroft's (2002, 2004 b) studies suggesting that 'semantically elaborative tasks', such as providing target words in an original context or focusing on usage contexts, may hinder the learning of formal aspects of L2 words. Barcroft did not recommend including a great number of these kinds of activities, especially in the initial stages of learning new words. The result of the current study, unlike Barcroft's view, indicated that for vocabulary learning (especially form recall), E of any type, which involves semantic elaboration, was more effective than S, even if S involves semantic elaboration. The results are in line with Laufer's (2001) suggestion for using vocabulary-focused activities that involve elaboration.

5.2.2.2 The Contribution of Search in Combination with Types of Evaluation

Investigating the contribution of S to vocabulary learning when it is combined with different types of E can also show the impact of search on the effectiveness of different types of evaluation. It was revealed that S could increase the effectiveness of R (moderate evaluation) for passive recall of words. However, it did not have any impact on the effectiveness of P, SW, and CW (varieties of strong evaluation) for vocabulary learning, which is in contrast with the ILH assumptions, given that by adding S to each type of E, the task involvement index goes higher and thus, the task effectiveness is supposed to increase. However, for P, SW, and CW, by adding S, the task effectiveness remained the same.

So far, it has been indicated that P, SW, and CW were more effective than R for vocabulary learning, except that P was similar to R for passive recognition of words. Putting this together with the above-mentioned conclusion that S could increase the effectiveness of R for passive recall does not show if the strength of R in the presence of S is comparable with the strength of P, SW, and CW. To investigate this, additional analyses were done and the

combination of R and S was compared with each of P, SW, and CW alone. The results indicated that first, S did not make R as effective as SW and CW for active recall and passive recognition. However, it did make R as effective as SW and CW for passive recall. This finding is not only consistent with the findings mentioned in the above paragraph that S could increase the effectiveness of R for passive recall, but also indicative that S could increase the effectiveness of R to the level comparable with the effectiveness of SW and CW for passive recall.

Second, S made R as effective as productive retrieval alone for all types of vocabulary learning measured in this study. As shown in the answer to the first research question, P was similar to R only for passive recognition of words. Adding S to R, as mentioned above, could make it as effective as P for all types of vocabulary learning, indicating that the differences between P and R (active and passive recall) faded away. Thus, it is concluded that S could increase the effectiveness of R for active and passive recall to the level comparable to the effectiveness of P.

These conclusions show that S plays the role of support for R, but not for other types of E. It may show that S has a differential impact on the effectiveness of different types of E and on different types of vocabulary learning. Thus, it is concluded that S does not carry a fixed involvement index and its contribution to vocabulary learning may depend on the type of evaluation it is combined with and the type of vocabulary learning. Although this goes against the ILH assumptions, Laufer and Hulstijn (2001), who proposed the ILH, were aware of some of the shortcomings of the framework. They pointed out that, although in determining the involvement index of a task they gave the same weight to the factors, the factors might not be similarly effective for vocabulary learning. The results indicated that the weight of S may be lower than E of any type because S could not even compete with R, which had the same involvement index for active recall. The higher weight of E compared to S was also advocated by Keating (2008), stating that E is a crucial factor for vocabulary learning.

The answer to the second research question can also reveal certain issues pertaining to the ILH. Based on the ILH assumptions, by adding the factors, the involvement index of any treatment and in turn, its effectiveness for vocabulary learning increases. However, the findings of this research indicated that this is not always the case. The results delineated that the involvement indices of treatments do not always reflect their effectiveness for vocabulary learning. For instance, although adding S to each type of evaluation increases the involvement index, it does not always increase the effectiveness of the treatment. In the case of P, SW, and CW, adding S increased the involvement index, but not the effectiveness of the combination. Therefore, by taking into account the results regarding the differential impact of S on different

types of E, calculating the involvement indices of the tasks is not as simple as before. For instance, comparing a task inducing moderate need (+), search (+) and R (+) with the involvement index of three and a task inducing moderate need (+), search (+) and P (++) with the involvement index of four does not show that the latter task is more effective because of the higher involvement index. The reason is that on the one hand, S can increase the effectiveness of R for active and passive recall to the level comparable to the effectiveness of P, but makes no meaningful contribution to the effectiveness of P. On the other hand, as the answer to the first research question indicated, P and R are similar in terms of their impact on passive recognition. Thus, it is likely that these two tasks are equally effective for vocabulary learning.

By putting together the conclusions drawn from the contribution of search alone to vocabulary learning and the contribution of search combined with other factors, the discussion regarding the second research question is summarized as follows:

1- The contribution of S is significantly lower than the contribution of SW, CW, and P to vocabulary learning.
2- The contribution of S is similar to the contribution of R to passive recall and recognition of words. However, for active recall, the contribution of S is significantly lower than that of R.
3- The contribution of S alone to vocabulary learning is marginal. Thus, on the continuum of the effectiveness for vocabulary learning that was mentioned above, where CW and SW are the strongest factors, S may have the weakest position.
4- S does not increase the effectiveness of SW, CW, and P for vocabulary learning.
5- S increases the effectiveness of R for passive recall to the level comparable with the effectiveness of SW and CW.
6- S increases the effectiveness of R for active and passive recall to the level comparable to the effectiveness of P.
7- The above two conclusions indicate that S plays the role of support for R.
8- S has a differential impact on the effectiveness of different types of E and on different types of vocabulary learning.
9- S does not carry a fixed involvement index and its effectiveness for vocabulary learning may depend on the type of evaluation it is combined with and the type of vocabulary learning.

By putting together the answers to the first and second research questions, it is possible to discuss Question 2a. As mentioned in 1.4, tasks can have the same involvement index, but different distribution of the factors, which means a different combination of the factors and their degrees. In their original article, Laufer and Hulstijn (2001) noted that the impact of the distribution of the factors was not clear. Later, studies provided conflicting evidence regarding the impact. For instance, Hu and Nassaji (2016) and Zou (2017) found out that tasks with the same involvement indices in which the factors were distributed differently were not equally effective for vocabulary learning. However, Kim (2011) and Rott (2007) supported the equal effectiveness of tasks with the same involvement indices and different distributions. There are also mixed findings regarding the impact on different types of vocabulary learning, namely receptive and productive (e.g., Bao, 2015).

The results of the present research provided mixed evidence, given that on several occasions, tasks with the same involvement index, but different distribution of the factors, were not equally effective for vocabulary learning. Examples are Groups 3-6 and 2-6 with the same involvement index of 3. Group 6 was more effective than Group 3 for all types of vocabulary learning and more effective than Group 2 for passive recall and recognition of words. However, Groups 6 and 2 were equally effective for active recall. It can be explained by the previous conclusion that task effectiveness for vocabulary learning does not always depend on the involvement indices. Type of evaluation and type of vocabulary learning also play their roles.

5.2.3 Research Question 3

3- What is the contribution of each degree of need to vocabulary learning?

As mentioned in the previous chapters, the contribution of each of dictionary search and evaluation to vocabulary learning could be compared by investigating the effect of each in the absence of the other. However, the contribution of search and evaluation could not be measured in the absence of need because need is a prerequisite for inducing search and evaluation and, unlike the other factors (search and evaluation), it could not be eliminated from the treatments. This had two consequences for this research. First, it was not possible to compare need with the other factors (search and evaluation of any type) in terms of their contribution to vocabulary learning. The reason is that the effect of need could be investigated in the absence of evaluation or search, but the effect of evaluation or search could not be investigated in the absence of need. Second, this study could only account for the differences that moderate and strong

degrees of need may make. Since in all learning situations, some degree of need is involved, investigating the impact of zero need or comparing it with the impact of moderate and strong degrees of need does not make sense.

To be able to account for the contribution of different degrees of need to vocabulary learning, there had to be two different tasks, one inducing a moderate degree of need and the other a strong degree. The tasks measuring the effectiveness of S and E had already induced moderate degrees of need. Thus, a specific task had to be designed to induce strong need and to be compared with previous tasks. This task (Section 2 of each lesson, Appendix 17, pp. 195 and 202) included reading passages containing some unknown words (the non-essential words) and tables that had to be completed using the information provided in the passages. Knowing the meanings of the non-essential words was not necessary for completing the tables. Decisions about learning these words were made by the learners themselves, not promoted by teachers or tasks. The participants were allowed to look up the words they wanted to know in a dictionary, but they had to circle the words they had looked up. This could help in identifying the words the participants chose to learn. For those participants who decided to learn the non-essential words, the task induced strong need and search. There was a reason behind allowing the participants to look up the words in a dictionary. If instead of asking the learners to look up the words, the meanings of words had been given, the participants would have been alerted to the focus on the words and in that case, the degree of need would not have been strong any more, given that under this condition, need had been imposed by the task.

After the task, learning the non-essential words (involving strong need, search, and some degrees of evaluation) had to be compared with learning the target words (involving moderate need, search, and some degrees of evaluation). Before making such comparisons, two issues had to be considered. First, the degree of evaluation for the non-essential words could not be determined. If the participants only looked up the meanings of words in the classroom, the degree of evaluation would be zero. However, if the participants practiced the words in speaking and writing or encountered them in listening and reading materials outside the research context, the degree of evaluation could be strong or moderate. The solution was to consider all degrees of evaluation for the non-essential words and compare them with the target words having the same degrees of evaluation. For instance, if the learning situation of the non-essential words induced strong need and search, it could be compared with the learning situation of the target words that induced moderate need and search. The comparison could differentiate between the impacts of strong and moderate degrees of need because the degrees of search and evaluation were the same (+ and -). In Table 5.1, which shows the comparisons,

N, S, and E refer to need, search, and evaluation. The degree of each factor was indicated above it. The superiority of each combination over the other was shown by < in each cell.

Table 5.1

Comparing the combinations of N, S, and E

Groups	Productive test (active recall)	Translation test (passive recall)	Matching test (passive recognition)
Non-essential words < target words learned by Group 9	$S + N^{++} < S + N^{+}$	$S + N^{++} < S + N^{+}$	$S + N^{++} < S + N^{+}$
Non-essential words < target words learned by Groups 1 and 3	$S + N^{++} + E^{+} < S + N^{+} + E^{+}$	$S + N^{++} + E^{+} < S + N^{+} + E^{+}$	$S + N^{++} + E^{+} < S + N^{+} + E^{+}$
Non-essential words < target words learned by Groups 2, 4, 5, 6, 7, and 8	$S + N^{++} + E^{++} < S + N^{+} + E^{++}$	$S + N^{++} + E^{++} < S + N^{+} + E^{++}$	$S + N^{++} + E^{++} < S + N^{+} + E^{++}$

Second, to achieve a result, either the performance of independent groups had to be compared or the performance of the same group on different tests had to be compared. However, in this study, the participants who learned the non-essential words were the same ones who learned the target words, but not everyone who learned the target words decided to learn the non-essential words. Therefore, the groups were neither independent nor the same and under such circumstances, the assumptions of statistical tests were violated. The solution was to calculate descriptive statistics.

The comparisons were made and it was shown that the mean scores of those who experienced strong need were lower than those who experienced moderate need. While it is not possible to make a strong claim based on descriptive statistics and state that strong need was significantly less effective than moderate need, the results certainly showed that strong need was not more effective than moderate need, which is contrary to the ILH assumptions.

In other words, in the instructional setting of the research, the superiority of strong need over moderate need faded away. The results are in conflict with the findings of the research by Lee and Pulido (2017), which is the only study that focused on the need factor of the ILH. In their research, the need factor was operationalized as topic interest and it was revealed that reading an interesting topic (higher degree of need) resulted in greater word learning compared

to reading topics of low interest (lower degree of need). However, Lee and Pulido's definition of need was different from the definition presented in the ILH, in other studies in the area (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2011), and in the present study, which is based on who imposes the need (task/teachers or the learners themselves) rather than topic interest. This fundamental difference may account for the conflicting results.

The explanation of the result may be that learners in the classroom context tend to be directed by their teachers and tasks. In such a setting, learners understandably give more importance to the words that are necessary for task completion and less importance to unnecessary words. This can be explained by Dornyei and Tseng's (2009, p. 119) account of the Motivational Task-processing System, which is made up of three mechanisms, namely task execution, task appraisal, and action control. Task execution in which learning occurs refers to learners' engagement in learning behaviours based on task instructions provided by the teacher or task designers. Task appraisal denotes the learners' appraisal of task completion by comparing their performance with the predicted one. Action control refers to the self-regulatory mechanism used by learners to enhance learning behaviour. In this study, the learners engaged in the given task based on the task instructions (task execution). Then, they completed the tasks with a low focus on the non-essential words, which were not necessary for task completion. Completing the task might satisfy the participants' appraisal of task achievement and thus they saw no point in focusing on the non-essential words to the same level that they focused on the target words.

It might be argued that the reason for the higher level of performance by the participants who learned the target words that involved moderate need, was that they received word-focused instruction and the type or degree of motivation (need) did not play a role. As mentioned in the above paragraphs, in the present study, the participants who learned the target words and thus experienced moderate need were provided with supplementary word-focused instruction, which might have provoked higher levels of processing compared to the participants who experienced strong need in the sense that they chose what to learn and look up and how to learn without receiving any word-focused instruction. A number of studies (e.g., Craik & Lockhart, 1972; Craik & Tulving, 1975; Hulstijn, 2001; Laufer & Hulstijn, 2001; Schmitt & Schmitt, 1995) have also found that word learning depends on processing information at a higher level or more elaborately, which according to Laufer (2001) is provoked by looking up words in a dictionary, using glosses, sentence completion, writing sentences, and writing compositions. Subsequent research supported her view by providing evidence that tasks with word-focused instruction are more effective for vocabulary learning compared to those without such

instruction (e.g., Nation, 2015; Sonbul & Schmitt, 2009; Webb, 2005; Wesche & Paribakht, 2000). Hulstijn (2001) also stated that vocabulary is learned through information processing, either at a higher level, which results in better retention, or at a lower level, which leads learners to process new words superficially or even to skip the words altogether. However, in this study, even in comparing those who experienced strong need and search and those who experienced moderate need and search, when the learning conditions were the same in terms of the type of elaboration (dictionary search) with the lowest involvement index, strong need was still not more effective than moderate need. It seems unlikely that elaboration or depth of processing could affect the result. The only possible difference between the groups was that for learning the target words, elaboration was instructed by the task, but for learning the non-essential words, elaboration originated from the learners' decision.

It can be concluded that in classroom contexts, when the goal is set by tasks or teachers, learners may not be highly subject to what is called strong need, which according to the ILH definition is imposed by the learners themselves. It is possible that strong need and moderate need (as defined by the ILH) are fundamentally different kinds of motivation, one of which (moderate need) is more relevant to a classroom setting when learners are focused on task completion and are directed by instructions and teachers, and the other with more relevance to self-directed learning.

Therefore, the results that strong need was not more effective than moderate need can be interpreted differently by redefining the need factor of the ILH and stating that the participants experiencing what is called moderate need underwent task-driven motivation for learning, which led them to learn by depending on task instructions and requirements. Nevertheless, the participants experiencing what is called strong need probably had self-driven motivation for learning and did not draw on it seriously because they were in the classroom setting and were completing a task based on instructions. They might have given more importance to the goals set by the tasks (in this case completing a table) than to their own personal learning goals (which could be vocabulary learning).

It may indicate that at least in a classroom setting, task-driven motivation for learning is more effective for vocabulary learning. Self-driven motivation for learning may be absent in tasks where the goals, instructions, and requirements are strictly determined by teachers or tasks. Thus, instead of having one motivational factor (need) with degrees, the involvement load may have two motivational factors, namely task-driven motivation for learning and self-driven motivation for learning.

This new definition of motivation is to some extent similar to the definition of motivation for vocabulary learning provided by the Technique Feature Analysis (TFA), as proposed by Nation and Webb (2011, p. 7) for analysing vocabulary task effectiveness. Instead of a narrow concept of motivation (need), the TFA asks three questions to give scores to the motivational factor of vocabulary-focused tasks:

1-Is there a clear vocabulary learning goal?

2-Does the activity motivate learning?

3-Do the learners select the words?

The first and second questions may refer to the task-driven motivation for learning and the third to self-driven motivation for learning. Each of the questions was given the score of one. Probably, in Nation and Webb's view, motivation created by the task (number one and two) with the total score of two made more contribution to the overall index of tasks compared to self-imposed motivation (number three) with a score of only one. It may show that Nation and Webb gave more prominence to the role played by motivation created by tasks than the role played by self-imposed motivation, at least in the instructional context. In this regard, their framework shows some support for the results of this study.

Chapter 6: Conclusion

6.1. Summary of Findings

This quantitative study was carried out to investigate the contribution of the involvement factors, namely need, search, and evaluation, to vocabulary learning. The data was collected from English learners in nine experimental groups and a control group. Each experimental group encountered new words through a task inducing a specific combination of the involvement factors. The results of the comparisons between the groups led to extending the evaluation factor, re-defining the need factor, and clarifying the role of the search factor.

The evaluation factor (E) was investigated by comparing four types of activities, namely composition-writing (CW), sentence-writing (SW), productive memory retrieval (P), and receptive memory retrieval (R). The first two activities have been investigated by previous studies and contradictory evidence has been found. The second two activities may have been compared previously, but not as different types of E inducing different degrees of involvement. Comparing four types of E indicated that CW and SW were similar in terms of their impacts on vocabulary learning and were the most powerful types of E. The less powerful type was P followed by R, which was found to be the weakest type of E. This conclusion extended the evaluation factor by proposing that instead of three degrees of prominence (-, +, and ++), E can be given four degrees: CW and SW induce very strong evaluation (+++). P and R induce strong (++) and moderate (+) degrees of evaluation respectively. In the absence of all these types, E is (-).

Regarding the search factor (S), two groups of conclusions were drawn depending on whether S appears alone or in combination with the types of E. In the first condition, where S appeared alone and its effectiveness was compared with the effectiveness of different types of E, it was shown that S was less effective than CW, SW, and P for vocabulary learning. S was also less effective than R for the most difficult kind of vocabulary learning (active recall of words). Only for easier types of learning (passive recall and recognition of words) was S similar to R. Thus, it was concluded that the contribution of S alone to vocabulary learning was marginal and weaker than any type of E.

In the second condition, it was indicated that S did not increase the effectiveness of CW, SW, and P for vocabulary learning. However, it increased the effectiveness of R for passive recall to the level comparable to the effectiveness of SW and CW and for active and passive recall to the level comparable to the effectiveness of P. Therefore, it was concluded

that, first, S played the role of support for R, but not for other types of E. Second, S had a differential impact on the effectiveness of different types of E and on different types of vocabulary learning. It may show that S does not carry a fixed involvement index, which is inconsistent with the Involvement Load Hypothesis (ILH) definitions in which each of the need, search, and evaluation factors was given degrees of prominence (involvement indices). The overall conclusion was that the contribution of S to vocabulary learning depends on the type of evaluation it is combined with and the type of vocabulary learning.

With respect to the need factor (N), the data indicated that the strong degree was not more effective than the moderate degree, which is also inconsistent with the ILH assumptions. The interpretation of the finding was that in classroom contexts, when tasks or teachers set the learning goals, learners might not be subject to strong need, which really should be imposed by learners themselves. This interpretation led to a conclusion that the need factor of the ILH should be re-defined. Instead of having one motivational factor (need) with three degrees of minus, moderate, and strong, the ILH may need to have two motivational factors, namely task-driven and self-driven motivation for learning. Moderate need with more relevance to classroom settings, when learners are focused on task completion based on task instructions and requirements, can be task-driven motivation. Strong need can be self-driven motivation for learning, which may not be drawn upon to any significant extent in classroom contexts, when learners give more importance to the goals set by tasks than to their own personal learning goals. Although this study could not indicate which of self-driven and task-driven motivation is more important for vocabulary learning, it did provide the basis for a proposal to extend (reformulate) the need factor.

6.2 Contributions of the Study

6.2.1 Contributions to Theory

This study contributes to vocabulary teaching theory in two main areas. First, investigating the contribution of the involvement factors could increase the potential of the ILH for operationalizing elaborate processing for L2 vocabulary learning. Extending the evaluation factor to give it more degrees (very strong, strong, moderate, and minus) may contribute to differentiating between a wider range of activities in terms of their impacts on vocabulary learning. Clarifying the role of search, alone and in combination with the types of evaluation, in learning different types of vocabulary may provide a more precise estimate of the power of

this factor in a variety of vocabulary-focused tasks. These suggested developments may increase the accuracy of the hypothesis for designing vocabulary-focused activities and judging their effectiveness. Re-defining the need factor, although did not affect the accuracy of the ILH in the way that extending the evaluation factor did, may enrich the hypothesis by providing a broader view of motivation through differentiating between self-driven and task-driven motivation for learning.

Second, investigating the contribution of the involvement factors could change some of the Involvement Load Hypothesis assumptions. With respect to E, it was found that retention of words is dependent not only upon the involvement load, but also on the characteristics of each type of evaluation. Characteristics such as using words in original contexts in CW and SW, production of form without using words in original contexts in P, and evaluation without production of form in R, each has its specific impact on learning different aspects of vocabulary. For instance, it was indicated that for developing active recall of words, production is a necessary element and any kind of evaluation that induces production can be effective, irrespective of the involvement load. For learning as many aspects of words as possible, using words in original contexts is required and the activities that induce it are the ones with the highest involvement load. For the easiest type of learning (passive recognition), evaluation may suffice and higher involvement load, production of form, and using words in original contexts may make no difference.

In connection with the above statements, it was concluded that task involvement indices do not always represent task effectiveness for all aspects of vocabulary learning. Task effectiveness may also depend on the type of evaluation, presence or absence of search, and the target type(s) of vocabulary. Accordingly, tasks with higher involvement indices are not always more effective than tasks with lower indices. In other words, increasing the involvement index of a task does not always increase its effectiveness for vocabulary learning. Moreover, tasks with the same involvement index, but different distribution of the factors may or may not be different in terms of their impacts on vocabulary learning.

These points do not discredit the ILH as a framework for designing and analysing vocabulary-focused tasks. They extend the evaluation factor, re-define the need factor, clarify the role of search factor, and bring the importance of type of vocabulary learning into focus when the hypothesis is used. Accordingly, designing, analysing, and adapting vocabulary-focused tasks are more complicated than before and the mentioned facets should be taken into account. For instance, based on the ILH predictions, task 1 presented in Table 6.1 is less effective than the other tasks because of its lower involvement load. Tasks 2 and 3 are equally

effective for vocabulary learning because of having the same involvement load. However, considering the modifications proposed in this research, predicting the effectiveness of these tasks may be more complicated. Their effectiveness may be summarized as follows:

Table 6.1

Comparing three tasks

Tasks				Task Involvement indices
1	Need +	Search +	Evaluation (receptive retrieval) +	3
2	Need +	Search +	Evaluation (productive retrieval) ++	4
3	Need +	Search +	Evaluation (composition-writing) ++	4

Tasks 1 and 2 are similar in terms of their impacts on active recall, passive recall, and passive recognition of words. The reason is that first, the answer to the first research question indicated that P and R are similar in terms of their impacts on passive recognition. Second, based on the answer to the second research question, while S does not change the effectiveness of P for vocabulary learning, it increases the effectiveness of R for active and passive recall to the level comparable to the effectiveness of P.

Tasks 1 and 3 are similar in terms of their impacts on passive recall of words, but task 3 is superior to task 1 for active recall and passive recognition. The reason is that first, the answer to the first research question indicated that CW is superior to R for active and passive recall and passive recognition. Second, based on the answer to the second research question, while S does not change the effectiveness of CW for vocabulary learning, it increases the effectiveness of R for passive recall to the level comparable to the effectiveness of CW.

Tasks 2 and 3, as the answer to the first research question indicated, are similar in terms of their impacts on active recall of words. However, when it comes to passive recall and recognition of words, task 3 is more effective than task 1. The answer to the second research question also indicated that S could not change the effectiveness of P and CW.

6.2.2 Contributions to Practice

This study contributes to vocabulary teaching practice in two main areas. Firstly, the findings can have implications for teachers and material writers, which are discussed in connection with Nation's view regarding vocabulary teaching. As Nation (2013, p. 95) points out, for vocabulary teaching, teachers should limit the time on each word and focus on the most

important aspects of vocabulary knowledge. The results of this study may contribute to designing activities that meet these requirements for vocabulary teaching. By excluding unnecessary involvement factors, less time-consuming and sufficiently effective vocabulary learning activities can be designed with the focus on learning specific aspects of vocabulary knowledge. The following paragraphs elaborate on the practical implications of the findings using sample activities.

This research indicated that dictionary search should be included in activities purposefully, considering the type of evaluation it is combined with and the required type of vocabulary learning to ensure dictionary search is worth the time given to it. Although Nation (2013, p. 426) states that dictionary look up itself is a vocabulary-focused task and previous studies (e.g., Knight, 1994; Luppescu & Day, 1993) indicate that it is effective for vocabulary learning, its marginal contribution to vocabulary learning and the time it takes may make it inferior to glosses in some situations, as the results of this study indicated. Dictionary search only increases the effectiveness of receptive memory retrieval for vocabulary learning. Thus, replacing dictionary search with glosses can be a good option for activities involving sentence and composition-writing and productive memory retrieval, so that time will be saved with no negative impact on vocabulary learning.

The type of evaluation should also be chosen by considering the time it takes and the target aspect of vocabulary. When developing active recall of words in a shorter time is required, an optimal activity can be one that induces productive memory retrieval without the need to include dictionary search and generative use of words. The reason is that first, as the answer to the first research question indicated, the types of evaluation that induce generative use (sentence and composition-writing) and productive memory retrieval are similarly effective for active recall of words. However, this research showed that productive memory retrieval takes a shorter time compared to sentence and composition-writing. Second, dictionary search, as the answer to the second research question indicated, does not increase the effectiveness of productive memory retrieval. This makes including dictionary search unnecessary. The shorter time needed for productive memory retrieval compared to sentence and composition-writing, and its similar degree of effectiveness for active recall, make it a good option for developing active recall of words in the limited class time. An activity designed for this purpose may provide exposure to target words accompanied by glosses to eliminate the need for dictionary search. This should be followed by a gap-filling activity in which learners have to retrieve the forms of the words they were exposed to from their memories in order to fill the gaps in the given sentences.

It is worth mentioning that in this study, for learning 12 to 13 words, the average time for productive memory retrieval was 17.5 minutes while the average times for sentence and composition-writing were 20 and 21.5 minutes respectively. Although the time difference between these activities is relatively small in a classroom context, for a larger number of words, for more difficult words, and for less proficient learners the gap might be larger. This leaves the decision to teachers to choose the best activity considering the time, the number and nature of target words, and their learners' ability. If the time difference between the activities is so insignificant that does not affect the class time, sentence and composition-writing are better options, given that the answer to the first research question showed that activities inducing generative use of words are more effective than other types of evaluation for learning as many aspects of words as possible. Since dictionary search does not affect the effectiveness of sentence and composition-writing for vocabulary learning, it can be excluded from the activities to save time. An activity designed for this purpose should provide exposure to target words accompanied by glosses. It should be followed by activities that encourage generative use of words, such as rewriting a story or written information gap tasks using the target words.

Another optimal activity with the purpose of developing active recall of words is one that induces receptive memory retrieval and dictionary search. As the answer to the second research question indicated, dictionary search increases the effectiveness of receptive memory retrieval for active recall of words to the level comparable to the effectiveness of productive retrieval. Thus, if a a group of learners have already gained the skill of dictionary search or if electronic dictionaries are used, receptive memory retrieval combined with dictionary search can be used for developing active recall of words. The activity designed for this purpose should expose learners to the target words accompanied by dictionary search. It should be followed by a gap-filling activity in which target words are given and learners have to retrieve the meanings of words from their memory in order to decide which word fits each of the gaps. This activity can also be helpful for developing passive recall of words, given that, based on the answer to the second research question, dictionary search increases the effectiveness of receptive retrieval for passive recall of words to the level comparable with the effectiveness of sentence and composition-writing. In this study, for learning 12 to 13 words, the average time for receptive memory retrieval was 14 minutes, which is shorter than the time given to sentence and composition-writing (20 and 21.5 minutes respectively). Thus, the combination of dictionary search and receptive retrieval can save the class time for learning active and passive recall of words, especially when electronic dictionaries are used, or larger number of words or difficult words are learned.

Secondly, the findings can also have implications for autonomous learners. In Nation's (2013) taxonomy of kinds of vocabulary learning strategies (Table 6.2), dictionary search and glossaries are categorized under the second class of strategies, namely sources, and are considered ways of consulting reference sources. Retrieving and generative use are categorized under the third class of strategies named processes. Thus, the impacts of these involvement factors, which vary in degrees, on learning different aspects of vocabulary can be brought to learners' attention as vocabulary learning strategies to let autonomous learners make informed decisions about activities that meet their personal vocabulary learning goals.

Table 6.2

Nation's (2013) taxonomy of types of vocabulary-learning strategies (p. 328)

General class of strategies	Types of strategies
Planning: choosing what to focus on and when to focus on it	Choosing words Choosing the aspects of word knowledge Choosing strategies Planning repetition and spending time
Sources: finding information about words	Analysing words Using context Consulting a reference source in L1 and L2 Using parallels in L1 and L2
Processes: establishing knowledge	Noticing Retrieving Generative use
Skill in use: enriching knowledge	Gaining in coping with input through listening and speaking Gaining in coping with output through reading and writing Developing fluency across the four skills

6.3 Limitations

Although attempts were made to control the variables, several limitations in the design of the study were identified. First, it may be argued that time on task could make a difference between the groups, in addition to the effects of the treatments. As mentioned in 3.4.3, time on task for the groups who experienced R, P, SW, and CW was approximately 14, 17.5, 20, and 21.5

minutes respectively. It indicates that at least the difference between R and the other groups and the difference between P and CW were considerable. Therefore, variable time on task, which was inherent in the study, given that production and generative use took more time, can be counted as an extraneous variable. However, Hulstijn and Laufer (2001) make the point that time on task is different from time on target items. As they argue, when activities take longer, the time devoted to the target words is not necessarily longer. Although learners may devote more time to writing a composition (time on task), they may spend a much shorter time on using target words (time on target items). Hill and Laufer (2003) provide an example to show the substantial difference between time on task and time on target items. As they state, a learner may spend fifty minutes on writing a composition, but the time given to checking the use of target words and incorporating them in the composition may be only fifteen minutes. Hulstijn and Laufer (2001) also argue that it does not make sense to separate time from task, and the time it takes is an essential part of each task. Their ILH, which is based on the Levels of Processing Theory, claims that the *quality* of learning activities matters not the *quantity* of work (time on task). By taking into account these arguments about time on task, it may be possible to state that the time difference between the activities in this study did not affect the results to any great extent.

Second, in this study, group homogeneity was judged by administering the X-Lex test, which is a passive recognition test estimating language learners' vocabulary size with a ceiling of 5000 words. The estimated vocabulary size, as Milton (2009, pp. 171-172) mentions, indicates learners' language proficiency based on the CEFR levels. The reason for choosing this test was the short administration time that could help in meeting the research goals under time pressure. It may be argued that this short test was not comprehensive enough to demonstrate the groups' homogeneity. Although this issue could be seen as one of the limitations of the study, it did not affect the results substantially. To ensure the participants' level of proficiency, the X-Lex test scores were double-checked with their previous performance scores (the American English File test administered by the language institutes). As a result, several participants were considered outliers and their data were excluded from the study.

Another limitation is related to the order of the vocabulary post-tests. An argument can be made that there may have been a learning effect from the productive test to the translation test. It is worth mentioning that while there could be some unexpected test order effects, attempts were made to limit such effects by taking into account the notion of degrees of

vocabulary knowledge (Laufer, Elder, Hill, & Congdon, 2004, pp. 206-207; Laufer & Goldstein, 2004, pp. 405-406) discussed in 3.2.4.3.1 to choose the best possible test order.

Moreover, in this study, it was not possible to fully control variables that could affect the need factor considering the fact that in Section 2 of both lessons, learners chose the words. A variety of reasons could affect learners' decisions to learn or not to learn the non-essential words included in this section.

Finally, in this study, dictionary search was operationalized based on Laufer and Hulstijn's (2001) definition of this involvement factor. In their definition, search is "the attempt to find the meaning of an unknown L2 word or trying to find the L2 word form expressing a concept (e.g., trying to find the L2 translation of an L1 word) by consulting a dictionary or another authority (e.g, a teacher)" (p. 14). However, in this definition, the processes that dictionary search entails, which may affect vocabulary learning, have not been taken into account. For instance, during dictionary search, learners need to compare senses of words that may induce moderate evaluation. In addition, learners have the opportunity to see example sentences provided by dictionaries. These processes could affect the vocabulary learning of participants who looked up the meanings of words in dictionaries compared to the groups who received glosses in which no example sentence was provided and no comparison of senses needed to be made. These processes could have created some differences between the groups and could be considered one of the limitations of this study. However, steps were taken to avoid the first process. The target and non-essential words chosen for the study were words whose meanings could be picked up immediately through dictionary search. The reason is that they were mostly words with only a few meanings and the tasks in which they appeared provided clear contexts. However, it is still possible that the participants were involved in making such comparisons. In spite of these limitations, the results of the study indicated that even if dictionary search could contribute to vocabulary learning through inducing evaluation and providing example sentences, it was still significantly less effective than the evaluation induced by any kind of memory retrieval and writing. Thus, even comparing senses of words and seeing words in example sentences could not bring about significant advantages for the search factor and could not affect the results of the study to any extent.

6.4. Suggestions for Future Research

The research findings and limitations may suggest further investigations of the involvement factors. First, as mentioned above, dictionary search may involve a variety of processes that

affect learning. This study indicated that even if dictionary search involves such processes, it is still significantly less effective than memory retrieval of any kind and writing. However, this study provided no evidence showing the extent to which these processes affect vocabulary learning or the degree of evaluation they induce, if any. Future research may be able to provide evidence.

Another finding was that as long as the learners' intention to know and use a word is affected by teacher instructions and task requirements, learners' personal goals in classroom settings might not have a considerable impact on vocabulary learning. That is why this study differentiated between task-driven and self-driven motivation for vocabulary learning and suggested including two need factors in the ILH (instead of having one need factor with three degrees). However, there is no evidence to indicate which of these two need factors is more effective for vocabulary learning and how degrees should be assigned to them. The reason is that the results were based on descriptive, not inferential, statistics. Future research may determine the weight of these two types of motivation for vocabulary learning.

Finally, in this study, need was operationalized based on Laufer and Hulstijn's (2001) definitions of this involvement factor. In their definition, need is learners' intention to learn or use a word. This definition does not cover other factors contributing to motivation such as topic interest, level of challenge, and interest built into an activity. Given that motivation may also depend on the mentioned factors and cannot be limited to learners' need, future research may compare the effectiveness of the mentioned factors and learners' intention. This may lead to developing the motivational factor of the ILH.

References

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, Del: International Reading Association.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge 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. https://doi.org/10.1016/j.system.2015.07.006

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*(2), 323-363. https://doi.org/10.1111/0023-8333.00186

Barcroft, J. (2004 a). Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals*, *37*(2), 200-208. https://doi.org/10.1111/j.1944-9720.2004.tb02193.x

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

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

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Boers, F., & Lindstromberg, S. (Eds.). (2008). *Cognitive linguistic approaches to teaching vocabulary and phraseology* (Vol. 6). Berlin, New York: Mouton de Gruyter.

Bruton, A., Lopez, M. G., & Mesa, R. E. (2011). Incidental L2 vocabulary learning: An impracticable term? *TESOL Quarterly*, *45*(4), 759-768. https://www.jstor.org/stable/41307665

Carr, N. T. (2011). *Designing and analysing language tests.* Oxford: Oxford University Press.

Carter, R. (1987). Vocabulary and second/foreign language teaching. *Language Teaching*, *20*(1), 3-16. https://doi.org/10.1017/S0261444800004225

Chacon-Beltrán, R., Abello-Contesse, C., & Torreblanca-López, M. D. M. (2010). Vocabulary teaching and learning: introduction and overview. In R. Chacon-Beltrán, C. Abello-Contesse, & M. D. M., Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 1-12). Bristol: Multilingual Matters.

Coady, J. (1997). L2 vocabulary acquisition: A synthesis of the research. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition*: *A rationale for pedagogy* (pp. 273-290). Cambridge: Cambridge University Press.

Cohen, A. D., & Aphek, E. (1980). Retention of second-language vocabulary overtime: Investigating the role of mnemonic associations. *System, 8*(3), 221-235. https://doi.org/10.1016/0346-251X(80)90004-4

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11*(6), 671-684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268. https://doi.org/10.1037/0096-3445.104.3.268

De Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning 56*, 3, 463–506. https://doi.org/10.1111/j.1467-9922.2006.00374.x

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1-56. https://doi.org/10.1111/0023-8333.00110

De La Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition, 24*(1), 81-112. https://doi.org/10.1017/S0272263102001043

Desrochers, A., & Begg, I. (1987). A theoretical account of encoding and retrieval processes in the use of imagery-based mnemonic. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes theories, individual differences, and applications* (pp. 56-77). New York: Springer.

Dornyei, Z., & Tseng, W. t. (2009). Motivational processing in interactional tasks. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction* (pp. 117-134). London: Routledge.

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. https://doi.org/10.1177/1362168811431377

Ellis, N., & Beaton, A. (1993 a). Psycholinguistic determinants of foreign language vocabulary

learning. *Language Learning 43*(4), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Ellis, N., & Beaton, A. (1993 b). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology Section A, 43*(4), 559-617. https://doi.org/10.1080/14640749308401062

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Ellis, R. (2009). Task-based language teaching: sorting out the misunderstandings. *International Journal of Applied Linguistics*, *9*(3), 221-246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Ellis, R. (2012). *Language teaching research and language pedagogy*. London: Wiley-Blackwell.

Ender, A. (2016). Implicit and explicit cognitive processes in incidental vocabulary acquisition. *Applied Linguistics*, *37*(4), 536–560. https://doi.org/10.1093/applin/amu051

Eysenck, M.W. (1982). Incidental learning and orienting tasks. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 197-228). New York: Academic Press.

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

Gonzalez-Fernandez, B., & Schmitt, N. (2017). Vocabulary acquisition. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 280-298). New York and London: Taylor & Francis.

Gray, P. (1999). *Psychology* (3rd ed.). New York: Worth Publishers.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research*, *11*(4), 459–479. https://doi.org/10.1177/1362168807080963

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. https://search-proquest-com.ezproxy.auckland.ac.nz/docview

Hu, H. C. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement Load Hypothesis versus Technique Feature Analysis. *System*, *56*, 28-39. https://doi.org/10.1016/j.system.2015.11.001

Huang, L. L., & Lin, C. C. (2014). Three approaches to glossing and their effects on vocabulary learning. *System, 44*, 127-136. https://doi.org/10.1016/j.system.2014.03.006

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. https://doi.org/10.1111/j.1540-4781.2012.01394.x

Hulstijn, J.H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258-286). Cambridge: Cambridge University Press.

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). USA, UK, Australia: Blackwell Publishing Ltd.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *SSLA, 27*, 129–140. https://doi.org/10.1017/S0272263105050084

Hulstijn, J. H. (2013). Incidental learning in second language acquisition. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 2632–7). Hoboken, NJ: John Wiley & Sons.

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

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51(3),* 539-558. https://doi.org/10.1111/0023-8333.00164

Jafari Gohar, M., Rahmanian, M., & Soleimani, H. (2018). Technique Feature Analysis or Involvement Load Hypothesis: estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859-869. https://www.ncbi.nlm.nih.gov/pubmed/29404913

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental Vocabulary acquisition? *Applied Linguistics, 19(3),* 357–377. https://doi.org/10.1093/applin/19.3.357

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365–386. https://doi.org/10.1177/1362168808089922

Kim, Y. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. In P. Robinson (Eds.), *Task-based language learning* (pp. 100-140). Malden, MA: Wiley-Blackwell.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal, 78*(3), 285-299. https://doi.org/10.1111/j.1540-4781.1994.tb02043.x

Koizumi, R. (2005). *Relationships between productive vocabulary knowledge and speaking performance of Japanese learners of English at the novice level* (Unpublished doctoral dissertation, University of Tsukuba, Japan).

Larson-Hall, J. (2010). A *guide to doing statistics in second language research using SPSS and R*. London: Routledge.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York: Routledge.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics, 19*(2), 255-271. https://doi.org/10.1093/applin/19.2.255

Laufer, B. (2001). Reading, word-focused activities and incidental vocabulary acquisition in a second language. *Prospect*, *16*(3), 44-54. http://www.ameprc.mq.edu.au/resources/prospect

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. https://doi.org/10.3138/cmlr.59.4.567

Laufer, B. (2005). Focus on Form in second language vocabulary Learning. *EUROSLA Yearbook, 5*(1), 223-250. https://doi.org/10.1075/eurosla.5.11lau

Laufer, B. (2010). Form-focused instruction in second language vocabulary learning. In R. Chacon-Beltrán, C. Abello-Contesse & M. D. M. Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 15-27). Bristol: Multilingual Matters.

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they

can be used for teaching purposes. *TESOL Quarterly, 47*(4), 867-872. https://www.jstor.org/stable/43267941

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202-226. https://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning & Technology, 3*(2), 58-76. https://eric.ed.gov/?id=ED462834

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language:The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1-26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163-176). London, New York: Routledge.

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*(4), 391-411. https://doi.org/10.1177/1362168811412019

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. https://doi.org/10.1177/1362168814559797

Lee, S., & Pulido, D. (2017). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research, 21*(1) 118–135. https://doi.org/10.1177/1362168816637381

Luppescu, S., & Day, R. R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning*, *43*(2), 263-279. https://doi.org/10.1111/j.1467-1770.1992.tb00717.x

Martinez-Fernandez, A. (2008). Revisiting the Involvement Load Hypothesis: awareness, type of task and type of item. In M. Bowles (Ed.), *Selected proceedings of the 2007*

second language research forum (pp. 210– 228). Somerville, MA: Cascadilla Proceedings Project.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analysing data: A model comparison perspective* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum associates, Inc.

Meara, P. (1990). A note on passive vocabulary. *Second Language Research*, *6*(2), 150-154. https://doi-org.ezproxy.auckland.ac.nz/10.1177/026765839000600204

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam, Philadelphia: John Benjamins Publication Company.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 84-102). Cambridge, New York: Cambridge University Press.

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

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspective on assessment and corpus analysis (pp.* 57-78). Emilia-Romagna, Italy: European Second Language Association. http://www.eurosla.org/monographs/EM02/TOC.pdf

Milton, J., & Fitzpatrick, T. (2014). *Dimensions of vocabulary knowledge* (Eds.). Basingstoke, Hampshire: Palgrave Macmillan.

Milton J., Wade, J. & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83-98). Bristol: Multilingual Matters.

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer, (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 79–100). Amsterdam: John Benjamins Publications.

Mondria, J. A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign Language. *Applied Linguistics, 12*(3), 249-267. https://doi.org/10.1093/applin/12.3.249

Nassaji, H., & Hu, H.-C. M. (2012). The relationship between task-induced involvement load

and learning new words from context. *IRAL*, *50*, 69-86. https://doi.org/10.1515/iral-2012-0003

Nation, I. S. P. (1993). Vocabulary size, growth, and use. In R. Schreuder and B Weltens (Eds.), *The bilingual lexicon* (pp. 115-134). Amsterdam, Philadelphia: John Benjamins Publication.

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques.* Boston, MA: Heinle.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*(1), 136-145. https://catalogue.library.auckland.ac.nz

Nation, I. S. P., & Webb, S. (2011). *Researching and analysing vocabulary.* Boston, MA: Heinle.

Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition*, *9*, 201-20. https://doi.org/10.1017/S0272263100000474

Palmer, H. E. (1921). *The principles of language study.* London: Harrap.

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*, 9–29. https://doi.org/10.18806/tesl.v11i1.623

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. 174-202). Cambridge: Cambridge University Press.

Phakiti, A. (2015). Quantitative research and analysis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics* (pp. 27-47). London, Oxford: Bloomsbury Academic.

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*,

11(1), 63-90. https://doi.org/10.1017/S027226310000783X

Piquer Piriz, A. M. (2008). Reasoning figuratively in early EFL: some implications for the development of vocabulary. In F. Boer & S. Lindstromberg (Eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology* (Vol. 6) (pp. 219-240). Berlin, New York: Mouton de Gruyter.

Pulido, D. (2009). How involved are American L2 learners of Spanish in lexical input processing tasks during reading? *Studies in Second Language Acquisition, 31*(1), 31-58. https://doi.org/10.1017/S0272263109090020

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437-447. https://doi.org/10.1016/j.jml.2009.01.004

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371. https://doi.org/10.1177/026553229301000308

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (209-227). Amsterdam, Philadelphia: John Benjamins.

Read, J. (2012). Piloting vocabulary tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (307-320). London, New York: Routledge.

Riazi, M. (2016). *The Routledge encyclopaedia of research methods in applied linguistics*. London: Routledge.

Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice*. New York: Cambridge university press.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*(s1), 1-36. https://doi.org/10.1111/j.1467-9922.2011.00641.x

Roediger, H. L., & Guynn, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 197–236). San Diego: Academic Press.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning, 57*(2), 165-199. https://doi.org/10.1111/j.1467-9922.2007.00406.x

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). New York: Cambridge University Press.

Schmitt, N. (2010 a). Key issues in teaching and learning vocabulary. In R. Chacon-Beltrán, C. Abello-Contesse & M. D. M. Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 28–40). Bristol, UK: Multilingual Matters.

Schmitt, N. (2010 b). *Researching vocabulary: A vocabulary research manual.* Basingstoke, UK: Palgrave Macmillan.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal, 49*(2), 133-143. https://doi.org/10.1093/elt/49.2.133

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913–951. https://doi.org/10.1111/lang.12077

Seligson, P., Latham-Koenig, C., & Oxenden, C. (2010). *American English file (2nd ed.).* Oxford: Oxford University Press.

Shintani, N. (2012). Input-based tasks and the acquisition of vocabulary and grammar: A process-product study. *Language Teaching Research, 16*(2), 253-279. https://doi.org/10.1177/1362168811431378

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics, 18,* 268-286. https://doi.org/10.1017/S0267190500003585

Snoder, P., & Reynolds, B. L. (2019). How dictogloss can facilitate collocation learning in ELT. *ELT Journal, 73*(1), 41-50. https://doi.org/10.1093/elt/ccy024

Sonbul, S., & Schmitt, N. (2009). Direct teaching of vocabulary after reading: is it worth the effort? *ELT Journal*, *64*(3), 253-260. https://doi.org/10.1093/elt/ccp059

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139-152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(04), 577-607.

https://doi.org/10.1017/S0272263109990039

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371-391. https://doi.org/10.1093/applin/16.3.371

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390-416. https://doi.org/10.1111/lang.12105

Van Patten, B. (1990). Attending to form and content in the input. *Studies in Second Language Acquisition*, *12*(3), 287-301. https://doi.org/10.1017/S0272263100009177

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics 22*(2), 217-234. https://doi.org/10.1017/S0142716401002041

Wallace, C. (2008). Vocabulary: The key to teaching English language learners to read. *The Education Digest, 73*(9), 36-39. https://catalogue.library.auckland.ac.nz

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33-52. https://doi.org/10.1017/S0272263105050023

Wesche, M. B., & Paribakht, T. S. (2000). Reading-based exercises in second language vocabulary learning: An introspective study. *The Modern Language Journal*, *84*(2), 196-213. https://doi.org/10.1111/0026-7902.00062

Zimmerman, C. B. (1997). Do reading and interactive vocabulary instruction make a difference? An empirical study. *TESOL Quarterly, 31*(1), 121-140. https://doi.org/10.2307/3587978

Zou, D. (2012). *A study of the components of the involvement load hypothesis: How involvement load should be allocated to "search" and "evaluation"* (Unpublished Ph.D. Dissertation. Hong Kong: City University of Hong Kong). http://lbms03.cityu.edu.hk/theses/abt/phd-en-b46907701a.pdf

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. https://doi.org/10.1177/1362168816652418

Appendices

Appendix 1: Lessons Used in the Pilot Study

Lesson 1: Extreme Weather

Section 1

Part A

What is the weather like in your country?

Please work in pairs and write two words that best explain the weather in each season in your country.

Spring	Summer	Fall	Winter

Part B

Which of the following conditions have happened in your country?

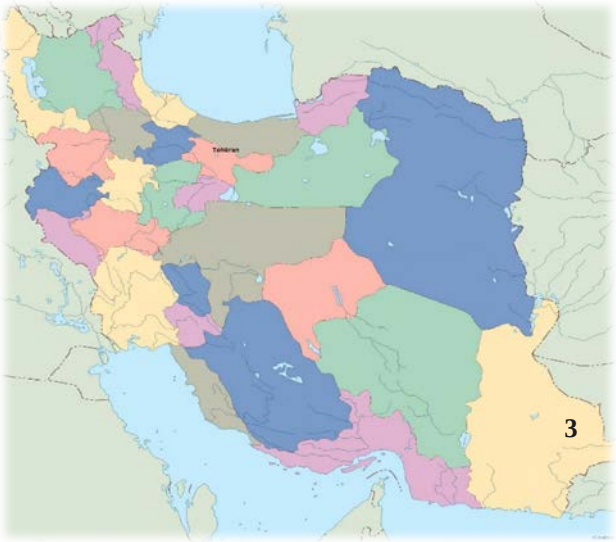
Please work individually. One of you circle the conditions that have occurred and the other circle the conditions that have not occurred. <u>For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.</u>

<u>For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.</u>

1- Devastating storm	2- Exceptional hot weather	3- Strong winds	4- Fall in sea levels
5- Gradual changes in weather	6- Low temperature	7- Lack of water	8- Poor agriculture

Now, please work in pairs and look at each other's answers. Does your friend agree with you?

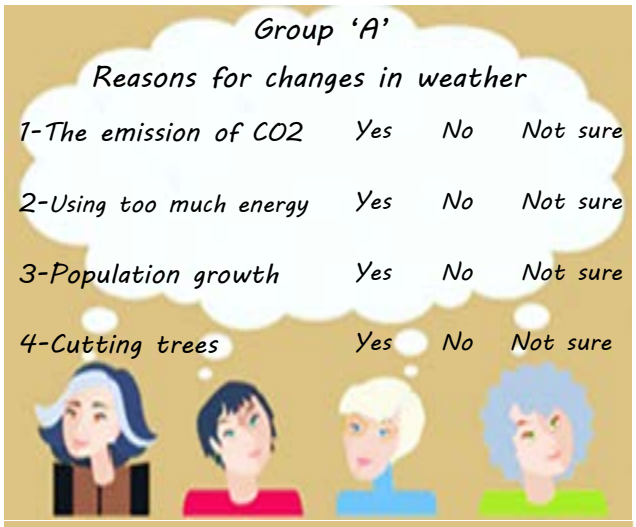
Please work together and write the number of each condition on the map to show where the condition has occurred (An example is number '3' on the map).



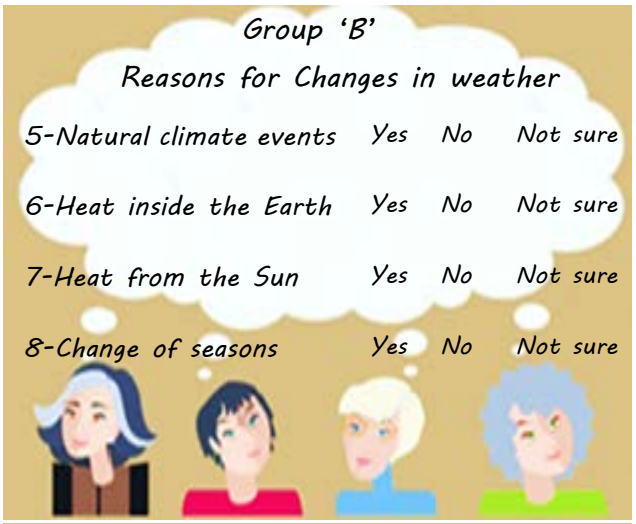
Devastating:	Causing a lot of damage	Agriculture:	Farming	Temperature:	A measure of how hot or cold a place or thing is
Exceptional:	Very unusual	Gradual:	Not sudden		

Part C

Please read the text individually and express your ideas by choosing yes, no, or not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.



The group's idea:
'Changes in weather originate from human activities'



The group's idea:
'Changes in weather are natural'

Now, which one do you agree with, A or B? Please work in pairs and compare your answers. If your friend does not agree with you go around the class and find someone who agrees with you. Then, put the eight reasons into three groups in the below table by writing the number of each reason in one of the rows:

Very important cause of changes in weather	
Important cause of changes in weather	
Not an important cause of changes in weather	

Originate: To happen or appear

Climate: The general weather condition in a particular area

Emission: The act of sending out light, heat, gas …

Part D

Now, you know your friend's ideas about weather change. Please one of you read the following sentences and express your ideas by choosing yes, no, not sure. The other one of you should guess his/her friend's ideas by choosing Agree, Disagree, or Not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.

For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

What can we do to stop changes in weather?			
1-It is not vital for people to use less gas and oil.	Agree	Disagree	Not sure
2-There is very little we can do to stop weather phenomena.	Agree	Disagree	Not sure
3-We must alert people to the problem.	Agree	Disagree	Not sure
4-We can only reduce the effects.	Agree	Disagree	Not sure
5-We must confront changes in weather in our country.	Agree	Disagree	Not sure
6-It is too late to stop global warming.	Agree	Disagree	Not sure
7-We should persuade people to change their habits.	Agree	Disagree	Not sure
8-We can conserve energy to stop heating of the Earth.	Agree	Disagree	Not sure

Now, please work in pairs and show your answers to each other. How many right guesses you have made?

Section 2

Please read the following text individually and list the factors that cause or do not cause heating of the Earth in the below table.

'Changes in weather are natural'

The Earth is affected by many natural events and one of them is weather change. Changes in weather have happened several times in the past and the Earth has often been warmer than now. Heating of the Earth is not caused by just one factor. One important factor during the past hundred years has been the Sun. The sun warms the Earth and a warmer planet will have more extreme weather such as storms, hot weather, and unusual rain. The earth's weather can also change because of cloud cover. Therefore, it is not true to say that human activities such as cutting trees, using too much energy, and population growth cause heating of the Earth. The reason for changes in weather is mostly natural. Then, there is very little we can do for heating of the Earth. We may control numerous factors such as population expansion and forest destruction, but we cannot control the blazing sun. The best thing we can do is to endeavour to reduce the effects of weather events. An example is warning people before storms that can rescue people's lives and helping victims after that.

Factors that cause heating of the Earth	Factors that do not cause heating of the Earth

Vital:	Necessary	Global:	Covering and affecting the whole world
Phenomena:	Unusual events	Persuade:	To make somebody decide to do something
Alert:	To warn someone about a problem or danger	Conserve:	To use a little of something
Confront:	To handle a problem		

Section 3

Receptive Memory Retrieval

Please work individually and fill each blank with one of the following words that is suitable (more than one word can be suitable for some of the sentences). Then, express your ideas about the sentences by circling yes, no or not sure.

1-There are many reasons for ……. change and human activity is just one of them.	Yes	No	Not sure
2- Pollution increases the ……. of the Earth and a warmer Earth will have more storms.	Yes	No	Not sure
3- Warmer weather and strong winds are not new ……. in the Earth's history. In the past, the Earth was warmer than now.	Yes	No	Not sure
4- Changes in weather are ……., so there is little danger in the near future.	Yes	No	Not sure
5- Sometimes, weather events are ……. and normal people cannot do much to stop these effects.	Yes	No	Not sure
6- One reason for ……. storms is air pollution.	Yes	No	Not sure
7- The ……. of CO2 can happen naturally in the environment and human activity is not the only reason for pollution.	Yes	No	Not sure
8- ……. problems are not specific to one country, so all the countries around the world are responsible for them.	Yes	No	Not sure
9- Warmer weather can make it difficult to grow plants so it can have harmful effects on ……. .	Yes	No	Not sure
10- The rains and storms we are experiencing may ……. from air pollution.	Yes	No	Not sure
11- To stop heating of the Earth, it is ……. for humans to change their habits.	Yes	No	Not sure
12- Extreme weather events such as strong winds ……. us to future dangers.	Yes	No	Not sure
13- It is necessary for all people around the words to ……. heating of the Earth.	Yes	No	Not sure
14- It is not easy to ……. people to plant trees or to use less water.	Yes	No	Not sure
15- By riding bikes and using less oil, people can ……. energy and stop heating of the Earth.	Yes	No	Not sure

Climate, Persuade, Temperature, Exceptional, Agriculture, Confront, Phenomena, Originate, Conserve, Gradual, Alert, Devastating, Vital, Global, Emission

Writing

What challenges other than heating of the Earth do you think people will face in the future?

Productive Memory Retrieval

Please work individually and fill in the blanks with suitable words. The first letter of each word is provided. Then, express your ideas about the sentences by circling yes, no or not sure.

1-There are many reasons for climate change and human activity is just one of them.	Yes	No	Not sure
2- Pollution increases the temperature of the Earth and a warmer Earth will have more storms.	Yes	No	Not sure
3- Warmer weather and strong winds are not new phenomena in the Earth's history. In the past, the Earth was warmer than now.	Yes	No	Not sure
4- Changes in weather are gradual, so there is little danger in the near future.	Yes	No	Not sure
5- Sometimes, weather events are devastating and normal people cannot do much to stop these effects.	Yes	No	Not sure
6- One reason for exceptional storms is air pollution.	Yes	No	Not sure
7- The emission of CO2 can happen naturally in the environment and human activity is not the only reason for pollution.	Yes	No	Not sure
8- Global problems are not specific to one country, so all the countries around the world are responsible for them.	Yes	No	Not sure
9- Warmer weather can make it difficult to grow plants so it can have harmful effects on agriculture.	Yes	No	Not sure
10- The rains and storms we are experiencing may originate from air pollution.	Yes	No	Not sure
11- To stop heating of the Earth, it is vital for humans to change their habits.	Yes	No	Not sure
12- Extreme weather events such as strong winds alert us to future dangers.	Yes	No	Not sure
13- It is necessary for all people around the words to confront heating of the Earth.	Yes	No	Not sure
14- It is not easy to persuade people to plant trees or to use less water.	Yes	No	Not sure
15-By riding bikes and using less oil, people can conserve energy and stop heating of the Earth.	Yes	No	Not sure

Writing

What challenges other than heating of the Earth do you think people will face in the future?

Sentence-writing

Please write a sentence for each word.

Climate
Persuade
Temperature
Exceptional
Agriculture
Confront
Phenomena
Originate
Conserve
Gradual
Alert
Devastating
Vital
Global
Emission

Composition-writing

Please write a short paragraph to say whether or not you agree with the fact that heating of the Earth is an important matter for the world.

You may find the following words helpful for writing the paragraph:

Climate, Phenomena, Gradual, Exceptional, Devastating, Vital, Alert, Emission, Originate, Global, Temperature, Agriculture, Confront, Conserve, Persuade

Lesson 2: The Internet

Section 1

Part A

What effect does the internet have on your life?

Please work in pairs and write three effects that using the internet has had on your life.

1-
2-
3-

Part B

Please read the following sentences individually and write the number of the sentences you agree with on a separate paper (for example: I agree with sentence number 1, 2, and 7). Please do not show your answers to each other. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.

For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

1-The majority of people use the internet for everyday life.
2-The internet has caused fundamental changes in human life.
3-Young people devote much time to the internet.
4-Children give their precious time to play online games.
5-The internet provides diverse kinds of fun for children.
6-The internet is an innovative system for learning.
7-Online programs decrease family members' time to talk to each other.

Now, please work in pairs. Without showing your answers, each one tells the number of the sentences she/he agrees with. The other student writes the numbers and guess his/her friends' view about the internet and tick in the below table. At the end, show what you ticked in that table to your friend. How many right guesses have you made?

My friend has a positive view about the internet	
My friend has a negative view about the internet	
My friend has neither a positive nor a negative view about the internet	

Majority:	Most of the people or things in a group	Diverse:	Very different from each other
Fundamental:	Basic	Innovative:	New
Devote:	To use all or most of your time and attention to do something	Decrease:	To make something smaller
Precious:	Valuable		

Part C

Please read the text individually and express your ideas by choosing yes, no, or not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

'A'

The internet is beneficial for children

The internet helps:

1-Social development	Yes	No	Not sure
2-Emotional development	Yes	No	Not sure
3-Language development	Yes	No	Not sure
4-Physical development such as working with fingers	Yes	No	Not sure

'B'

The internet is harmful for children

The internet causes:

5-Isolation from society	Yes	No	Not sure
6-Neck and head pain	Yes	No	Not sure
7-Addiction to computer games	Yes	No	Not sure
8-Deficiency of physical exercise	Yes	No	Not sure

Now, which one do you agree with, A or B? Please work in pairs and compare your answers. If your partner does not agree with you go around the class and find someone who agrees with you. Then, put the eight effects into the three groups in the below table by writing the number of each effect in one of the rows

This usually happens when children use the internet	
This may or may not happen when children use the internet	
This never happens when children use the internet	

Beneficial:	Having a helpful or useful effect	Addiction:	Being unable to stop doing something as a habit
Isolation:	When a person or thing is alone and separate from others	Deficiency:	Lack of something

Part D

Now, you know your friend's ideas about the internet. Please one of you read the following sentences and express your ideas by choosing yes, no, not sure. The other one of you should guess his/her friend's ideas by choosing Agree, Disagree, or Not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

What should and should not be done?			
1-Parents should restrict children's access to the internet.	Agree	Disagree	Not sure
2- Parents should permit children to use the internet only for doing homework.	Agree	Disagree	Not sure
3-Parents should enforce their rules for children's access to the Internet.	Agree	Disagree	Not sure
4-Parents should help children to use the internet for both learning and fun.	Agree	Disagree	Not sure
5- Parents should do nothing to stop children from using the internet.	Agree	Disagree	Not sure
6-Teachers should use the internet for its great impact on learning at schools.	Agree	Disagree	Not sure

Now, please work in pairs and show your answers to each other. How many right guesses you have made?

Section 2

Please read the following text individually and list old and new hobbies in the below table.

'Using the internet is harmful'

Only fifty years ago, life was very different. Young people had enough time to be with their families, go out for some exercise, or enjoying a book. But, these days, the internet has caused changes in everyone's life. For young people, the internet is a system that provides different types of entertainment such as games, movies, and online programs. But, parents think children's illogical use of the internet is irritating and not useful at all. Children give much of their time to use the internet with no specific aim. They rarely have time to play with friends and may never experience intimacy. This may limit their chance to have healthy relationships. Children may also spend their sleep and meal times on online activities. They don't go out to participate in play and this lack of physical activity can be very harmful. To reduce the effects of the internet on children's life and health, in all families, there should be some rules to stop internet overuse. Parents should help children to acquire healthy habits for computer use.

Hobbies for fifty years ago	Internet hobbies

Restrict:	To limit something	Permit:	To allow somebody to do something
Enforce:	To make sure people do what they have to do	Impact:	The effect that something has

Section 3

Receptive Memory Retrieval

Please work individually and fill each blank with one of the following words that is suitable (more than one word can be suitable for some of the sentences). Then, express your ideas about the sentences by circling yes, no or not sure.

1- People cannot trust the internet because the ……. of online information and news are false.	Yes	No	Not sure
2- There is a ……. difference between people's life today and their life in the past because of the internet.	Yes	No	Not sure
3- If children ……. time to the internet they won't have time to study and play with friends.	Yes	No	Not sure
4-It is possible for people to control the ……. of the internet on their lives.	Yes	No	Not sure
5- Social ……. happens when people spend their time on the internet and don't meet family and friends.	Yes	No	Not sure
6- Like any type of ……., when people use the internet too much, they cannot stop it and do their normal life activities.	Yes	No	Not sure
7- People who use the internet for long hours may be in danger of a ……. of physical activity.	Yes	No	Not sure
8- Using the internet can ……. the time people spend for daily activities.	Yes	No	Not sure
9- In the modern world, it is not possible for parents to ……. rules for not using the internet.	Yes	No	Not sure
10- Parents and schools should ……. children to use the internet only for learning.	Yes	No	Not sure
11- Without rules for using the internet, it cannot be ……. for children.	Yes	No	Not sure
12- The internet is interesting for children because it provides ……. types of online games.	Yes	No	Not sure
13- Parents should help children to ……. the hours of playing with computers to have time for other activities such as playing outside.	Yes	No	Not sure
14- Schools can use the internet as an ……. system for learning.	Yes	No	Not sure
15-The internet is a ……. tool for children's development if they use it in a suitable way.	Yes	No	Not sure

Permit, Fundamental, Addiction, Precious, Majority, Deficiency, Innovative, Devote, Decrease, Diverse, Impact, Beneficial, Isolation, Restrict, Enforce

Writing

Life has changed a lot in the last fifty years. What changes other than the internet do you notice?

Productive Memory Retrieval

Please work individually and fill in the blanks with suitable words. The first letter of each word is provided. Then, express your ideas about the sentences by circling yes, no or not sure.

1- People cannot trust the internet because the majority of online information and news are false.	Yes	No	Not sure
2- There is a fundamental difference between people's life today and their life in the past because of the internet.	Yes	No	Not sure
3- If children devote time to the internet they won't have time to study and play with friends.	Yes	No	Not sure
4-It is possible for people to control the impact of the internet on their lives.	Yes	No	Not sure
5- Social isolation happens when people spend their time on the internet and don't meet family and friends.	Yes	No	Not sure
6- Like any type of addiction, when people use the internet too much, they cannot stop it and do their normal life activities.	Yes	No	Not sure
7- People who use the internet for long hours may be in danger of a deficiency of physical activity.	Yes	No	Not sure
8- Using the internet can restrict the time people spend for daily activities.	Yes	No	Not sure
9- In the modern world, it is not possible for parents to enforce rules for not using the internet.	Yes	No	Not sure
10- Parents and schools should permit children to use the internet only for learning.	Yes	No	Not sure
11- Without rules for using the internet, it cannot be beneficial for children.	Yes	No	Not sure
12- The internet is interesting for children because it provides diverse types of online games.	Yes	No	Not sure
13- Parents should help children to decrease the hours of playing with computers to have time for other activities such as playing outside.	Yes	No	Not sure
14- Schools can use the internet as an innovative system for learning.	Yes	No	Not sure
15-The internet is a precious tool for children's development if they use it in a suitable way.	Yes	No	Not sure

Writing

Life has changed a lot in the last fifty years. What changes other than the internet do you notice?

Sentence-writing

Please write a sentence for each word.

Permit
Fundamental
Addiction
Precious
Majority
Deficiency
Innovative
Devote
Decrease
Diverse
Impact
Beneficial
Isolation
Restrict
Enforce

Composition-writing

Please write a short paragraph to express your idea about the effects of using the internet on people's lives.

You may find the following words helpful for writing the paragraph:

Majority, Fundamental, Innovative, Diverse, Precious, Beneficial, Devote, Isolation, Restrict, Impact, Addiction, Deficiency, Decrease, Enforce, Permit

Appendix 2: Vocabulary Pre-test

Student's name…..... Classroom number ……

Please write the meaning of each word in **Farsi**.

1	Natural		31	Alert	
2	Development		32	Provide	
3	Climate		33	Diverse	
4	Temperature		34	Precious	
5	Agriculture		35	Devote	
6	Reason		36	Harmful	
7	Endeavour		37	Numerous	
8	Global		38	Permit	
9	Exceptional		39	Population	
10	Vital		40	Decrease	
11	Devastating		41	Enforce	
12	Society		42	Fundamental	
13	Originate		43	Phenomena	
14	Healthy		44	Enough	
15	Confront		45	Emission	
16	Impolite		46	Connect	
17	Persuade		47	Beneficial	
18	Majority		48	Restrict	
19	Impact		49	Reduce	
20	Explain		51	Isolation	
21	Addiction		51	Gradual	
22	Illogical		52	Acquire	
23	Deficiency		53	Conserve	
24	Habit		54	Innovative	
25	Participate		55	Irritating	
26	Destruction		56	Blazing	
27	Damage		57	Promise	
28	Rescue		58	Intimacy	
29	Magazine		59	Experience	
30	Aim		60	Expansion	

Appendix 3: Productive Vocabulary Test for the Pilot Study

For Lesson 1

Student's name…..... Classroom number ……

Please write one English word for each description. The first letter of each word is provided for you.

	Explanations	Words
1	The general weather condition in a place	Climate
2	Unusual events	Phenomena
3	Farming	Agriculture
4	Damage to something	Destruction
5	A measure of how hot or cold a place or thing is	Temperature
6	The production and sending out of gas	Emission
7	Growth	Expansion
8	Happening slowly over a long time	Gradual
9	For the whole world	Global
10	Causing a lot of harm	Devastating
11	Very hot	Blazing
12	Many	Numerous
13	Very unusual	Exceptional
14	Very necessary	Vital
15	To make somebody aware of something	Alert
16	To happen or appear	Originate
17	To use a little of something and keep it for a long time	Conserve
18	To save somebody or something from danger	Rescue
19	To encourage somebody to do something by giving good reasons	Persuade
20	To handle a problem	Confront
21	To try very hard to do something	Endeavour

For Lesson 2

Student's name…..... Classroom number ……

Please write one English word for each description. The first letter of each word is provided for you.

	Explanations	Words
1	being alone or separate from others	Isolation
2	the largest number of people or things	Majority
3	the effect that something has	Impact
4	the reason for doing something	Aim
5	not having enough of something	Deficiency
6	being unable to stop doing something	Addiction
7	having a close relationship	Intimacy
8	very helpful	Beneficial
9	very different from each other	Diverse
10	something that makes others angry	Irritating
11	not reasonable	Illogical
12	new	Innovative
13	basic	Fundamental
14	valuable	Precious
15	to limit something	Restrict
16	to take part in an activity	Participate
17	to give time and attention to something	Devote
18	to gain something	Acquire
19	to make sure people do what they have to do	Enforce
20	to allow somebody do something	Permit
21	to make something smaller	Decrease

Appendix 4: Translation Test for the Pilot Study

For Lesson 1

Student's name…..... Classroom number ……

It the below table, please write the meaning of each word in **Farsi**. For each word provide only one meaning that you think is the best.

1	Phenomena		12	Originate	
2	Emission		13	Conserve	
3	Climate		14	Confront	
4	Temperature		15	Persuade	
5	Agriculture		16	Rescue	
6	Gradual		17	Destruction	
7	Global		18	Expansion	
8	Exceptional		19	Endeavour	
9	Vital		20	Blazing	
10	Devastating		21	Numerous	
11	Alert				

For Lesson 2

Student's name…..... Classroom number ……

It the below table, please write the meaning of each word in **Farsi**. For each word provide only one meaning that you think is the best.

1	Deficiency		12	Addiction	
2	Fundamental		13	Isolation	
3	Innovative		14	Impact	
4	Beneficial		15	Majority	
5	Diverse		16	Intimacy	
6	Precious		17	Irritating	
7	Devote		18	Participate	
8	Restrict		19	Aim	
9	Permit		20	Logical	
10	Enforce		21	Acquire	
11	Decrease				

Appendix 5: Matching Test for the Pilot Study

For Lesson 1

Student's name…..... Classroom number ……

Please match each word on the left with only one meaning on the right.

1- Climate H
2-Phenomena E
3-Agriculture B
4-Destruction F

A. fishing
B. farming
C. safe condition
D. strange stories
E. unusual events
F. damage to something
G. making something new
H. general weather condition

5- Temperature F
6- Emission D
7- Expansion A

A. growth
B. health
C. the production of energy
D. the production and sending out of gas
E. a unit that shows how wet or dry a place is
F. a measure of how hot or cold a place or thing is

8-Gradual A
9-Global E
10-Devastating D
11-Blazing H

A. happening slowly over a long time
B. happening in a specific place
C. causing a lot of questions
D. causing a lot of harm
E. for the whole world
F. for local people
G. very angry
H. very hot

12- Numerous B
13- Exceptional D
14- Vital F

A. Most
B. Many
C. very heavy
D. very unusual
E. very delicious
F. very necessary

15- Alert C

16- Originate E

17- Conserve A

A. to use a little of something and keep it for a long time
B. to spend a long time for doing something and do it completely
C. to make somebody aware of something
D. to make somebody believe something
E. to happen or appear
F. to find something

18-Rescue B

19- Persuade H

20- Confront C

21- Endeavour E

A. to make things easier
B. to save somebody form danger
C. to handle a problem
D. to put something out of sight
E. to try very hard to do something
F. to think carefully about something
G. to find detailed information about something
H. to make somebody do something by giving good reasons

For Lesson 2

Student's name…..... Classroom number ……

Please match each word on the left with only one meaning on the right.

1-Isolation G
2- Majority H
3- Impact E
4-Aim F

A. being depressed
B. the ability to do something
C. a group of people or things
D. the value that something has
E. the effect that something has
F. the reason for doing something
G. being alone or separate from others
H. the largest number of people or things

5- Deficiency D

6-Addiction F

7-Intimacy B

A. having a problem
B. having a close relationship
C. not having access to technology
D. not having enough of something
E. being able to do necessary things
F. being unable to stop doing something

8- Beneficial B

9- Diverse F

10- Irritating G

11-Illogical D

A. very usual
B. very helpful
C. not believable
D. not reasonable
E. unsuitable use of technology
F. very different from each other
G. something that makes others angry
H. something that makes people laugh

12- Innovative A
13- Fundamental B
14- Precious D

A. new
B. basic
C. rapid
D. valuable
E. possible
F. difficult

15- Restrict G

16- Participate E

17- Devote C
18-Acquire H

A. to find detailed information about something
B. to give less value and importance to something
C. to give time and attention to something
D. to make sure something is right
E. to take part in an activity
F. to question something
G. to limit something
H. to gain something

19- Enforce A

20- Permit D
21- Decrease E

A. to make sure people do what they have to do
B. to let somebody tell you what to do
C. to make sure something is enough
D. to allow somebody do something
E. to make something smaller
F. to ask for information

Appendix 6: Ethics Approval for the Pilot Study

Research Office
Post-Award Support Services

The University of Auckland
Private Bag 92019
Auckland, New Zealand

Level 10, 49 Symonds Street
Telephone: 64 9 373 7599
Extension: 83711
Facsimile: 64 9 373 7432
ro-ethics@auckland.ac.nz

UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE (UAHPEC)

19-Sep-2017

MEMORANDUM TO:

Dr Rosemary Erlam
App Lang Studies & Linguistics

Re: Application for Ethics Approval (Our Ref. 019930): Approved with comment

The Committee considered your application for ethics approval for your study entitled **The Impact of Task Type on Language Learning**.

Ethics approval was given for a period of three years with the following comment(s):

1. Please ensure that your translated public documentation for students are forwarded to ro-ethics@auckland.ac.nz.
2. Please ensure that you reformat all other public documentation.

The expiry date for this approval is 19-Sep-2020.

If the project changes significantly you are required to resubmit a new application to UAHPEC for further consideration.

If you have obtained funding other than from UniServices, send a copy of this approval letter to the Activations team in the Research Office, at ro-awards@auckland.ac.nz. For UniServices contracts, send a copy of the approval letter to the Contract Manager, UniServices.

The Chair and the members of UAHPEC would be happy to discuss general matters relating to ethics approvals if you wish to do so. Contact should be made through the UAHPEC Ethics Administrators at ro-ethics@auckland.ac.nz in the first instance.

Please quote Protocol number **019930** on all communication with the UAHPEC regarding this application.

(This is a computer generated letter. No signature required.)

UAHPEC Administrators
University of Auckland Human Participants Ethics Committee

c.c. Head of Department / School, App Lang Studies & Linguistics
Mrs Mandana Hazrat
Mr Keith Montgomery

Additional information:

1. Do not forget to fill in the 'approval wording' on the Participant Information Sheets, Consent Forms and/or advertisements, giving the dates of approval and the reference number. This needs to be completed, before you use them or send them out to your participants.

2. At the end of three years, or if the study is completed before the expiry, you are requested to advise the Committee of its completion.

3. Should you require an extension or need to make any changes to the project, please complete the online Amendment Request form associated with this approval number giving full details along with revised documentation. If requested before the current approval expires, an extension may be granted for a further three years, after which time you must submit a new application.

Appendix 7: Ethics Approval for the Main Study

Research Office
Post-Award Support Services

The University of Auckland
Private Bag 92019
Auckland, New Zealand

Level 10, 49 Symonds Street
Telephone: 64 9 373 7599
Extension: 83711
Facsimile: 64 9 373 7432
ro-ethics@auckland.ac.nz

UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE (UAHPEC)

11-Jun-2018

MEMORANDUM TO:

Prof John Read
App Lang Studies & Linguistics

Re: Application for Ethics Approval (Our Ref. 021365): Approved with comment

The Committee considered your application for ethics approval for your study entitled **The Impact of Task Type on Language Learning**.

Ethics approval was given for a period of three years with the following comment(s):

1. Thank you for your response to the previous comments. As the earlier application (019930) concerned a pilot and this is the main study, it is appropriate to treat it as a new application (though connected to 019930) and not as an amendment.

The expiry date for this approval is 11-Jun-2021.

If the project changes significantly you are required to resubmit a new application to UAHPEC for further consideration.

If you have obtained funding other than from UniServices, send a copy of this approval letter to the Activations team in the Research Office, at ro-awards@auckland.ac.nz. For UniServices contracts, send a copy of the approval letter to the Contract Manager, UniServices.

The Chair and the members of UAHPEC would be happy to discuss general matters relating to ethics approvals if you wish to do so. Contact should be made through the UAHPEC Ethics Administrators at ro-ethics@auckland.ac.nz in the first instance.

Please quote Protocol number **021365** on all communication with the UAHPEC regarding this application.

(This is a computer generated letter. No signature required.)

UAHPEC Administrators
University of Auckland Human Participants Ethics Committee

c.c. Head of Department / School, App Lang Studies & Linguistics
Mrs Mandana Hazrat
Mr Keith Montgomery

Additional information:

1. Do not forget to fill in the 'approval wording' on the Participant Information Sheets, Consent Forms and/or advertisements, giving the dates of approval and the reference number. This needs to be completed, before you use them or send them out to your participants.

2. At the end of three years, or if the study is completed before the expiry, you are requested to advise the Committee of its completion.

3. Should you require an extension or need to make any changes to the project, please complete the online Amendment Request form associated with this approval number giving full details along with revised documentation. If requested before the current approval expires, an extension may be granted for a further three years, after which time you must submit a new application.

Appendix 8: A Sample of Students' Participation Information Sheet for the Pilot Study

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

Participant Information Sheet

To: Student

Project title: The Impact of Task Type on Language Learning

Researcher: Mandana Hazrat
Supervisor: Dr. Rosemary Erlam

I am a doctoral candidate in Applied Language Studies at the University of Auckland in New Zealand. I am interested in finding out what type of classroom activities best help students learn English. I am inviting you to be part of a project that will help me find this out.

I have chosen your class because you are at an intermediate level of proficiency and are familiar with completing pair activities in your language lessons.

As part of this project, you will receive two lessons taught by your teacher and you will complete some tests. The lessons are designed for students at your level and your teacher has agreed that these lessons and tests are suitable for your class. Before the lessons, you will first take two tests in one session that will take about 30 minutes. In the second session, you will receive Lesson 1 that will take about one and a half hours. In the third session, you will complete three tests that will take about an hour. The procedures of the second and third sessions will be repeated for Lesson 2 except that you will also complete a questionnaire about the lessons in the last session that may take about 15 minutes. Therefore, the whole study will take about five hours and 45 minutes over five sessions.

I am asking you to allow me to use your test and questionnaire results for my study. I would also like you to allow me to have access to information about your previous class grades from the language institute. If you do not agree to participate in the study by allowing me to have access to your test and questionnaire results and grade information, you will still complete all the tests and lessons as part of your normal instruction. However, your test and questionnaire results will not be used in the research. It is likely that the experience of completing the tests and lessons will help you improve your English. If you agree to participate in the study, you may help me know how to help students like you to learn English.

It is up to you to agree to this and if you agree, you can change your mind before 15th December 2017 without the need to give a reason. If you agree to participate in the research, please fill out the Consent Form at home. In the next session, whether you fill the Consent Form or not, please return it to the researcher. If you decide to change your mind later, just text the researcher or give her a phone call. Then, you will complete all the tests and lessons, but

the results of your tests will not be used in this research project. Participation or non-participation in this project will not affect your relationship with your teacher or the Head of the Language Institute.

All information about you will be kept by the researcher in a locked cabinet and on a password protected computer both in Iran and in New Zealand and will be shredded or deleted after six years. If the information you provide is published, this will be done in a way that no one can identify you, your teacher, or your school. If you are interested to know about the result of the research, you can provide your email or postal address on the Consent Form. I will send you a copy of any publication that results from the study.

All participants in your class who agree to take part in the study by allowing me to have access to their test and questionnaire results and grade information will be offered the chance to enter a draw to win a book voucher to the value of $50 NZD. If you are one of the students who agree to participate and would like to enter the draw, please note this on your Consent Form and provide your phone number. At the end of the research, the researcher will write the names on pieces of papers in privacy and for each class, one name will be drawn randomly as the winner. The winners will be informed by the researcher though phone call.

If you are happy to participate, please fill out the Consent Form and return it to the researcher. You can discuss any issue with the researcher at any time by email or phone call or you are welcome to contact the following at the University of Auckland:

Researcher	Supervisor and Senior Lecturer	Head of School of Cultures, Languages and Linguistics
Mandana Hazrat Applied Language Studies and Linguistics University of Auckland Arts 2 - Bldg 207 Level 3, Room 313 18 SYMONDS ST AUCKLAND 1010 New Zealand Iranian Phone number: +09123212684 Email: mhaz902@aucklanduni.ac.nz	Dr. Rosemary Erlam Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 2, Room 212 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 7081 Email: r.erlam@auckland.ac.nz	Professor Gary Barkhuizen Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 3, Room 318 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 8197 Email: g.barkhuizen@auckland.ac.nz

For any concerns regarding ethical issues you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Research Office, Private Bag 92019, Auckland 1142. Telephone: 09 373 7599 ext. 83711. Email: ro-ethics@auckland.ac.nz

Approved by the University of Auckland Human Participants Ethics Committee on 19/09/2017 for three years, Reference Number 019930

Appendix 9: A Sample of Teachers' Participant Information Sheet for the Main Study, Experimental Groups

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

ARTS

Participant Information Sheet

To: Teachers

Project title: The Impact of Task Type on Language Learning

Researcher: Mandana Hazrat
Supervisor: Prof. John Read

I am a doctoral candidate in Applied Language Studies at the University of Auckland in New Zealand. I am interested in finding out what type of classroom activities best help students learn English. I am inviting you to be part of a project that will help me find this out. I have chosen your students because they are at an intermediate level of proficiency and are familiar with completing pair activities.

This is an intervention study in which the students will receive two lessons designed by the researcher for learners at an intermediate level to be used as additional materials. Before the lessons, the students will first take two tests in one session that will take about 30 minutes. In the second session, they will receive Lesson 1 that will take about one and a half hours. In the third session, the students will complete three tests that will take about 45 minutes. The procedures of the second and third sessions will be repeated for Lesson 2. In the last session of the study that will be held two weeks later, the tests of Lesson 1 and two will be administered as delayed post-tests. This session will take about one and a half hours. Therefore, the whole study will take about six hours over six sessions.

I am asking you to allow me to carry out the research in your class. If you agree, I will provide all the teaching and testing materials. I will ask you to check that the lessons are suitable for your class. I will then ask you to teach the lessons and administer the tests. Before each session, I will let you know how to carry out the procedures and how to use the teaching and testing materials. I will also attend the teaching and testing sessions. If you agree to participate in the study, you may help me know how to better help teachers like you to design effective lessons.

Participation in this project will be voluntary and you can withdraw before 30th of August 2018 without the need to give a reason. Your participation or non-participation will not in any way affect your relationship with the Head of the Language Institute. The students' participation is also voluntary and they can withdraw before 30th of August 2018 without providing a reason. The Head of the Language institute has given the assurance about participation or non-participation.

I will ask you to distribute the Participant Information Sheets and CF among the students and I will be in the classroom to answer the students' questions. The students who

agree to participate will fill out the Consent forms at home and return them to the researcher in the next session.

All information about the participants will be kept by the researcher in a locked cabinet and on a password protected computer both in Iran and in New Zealand and will be shredded or deleted after six years. If the information about the research is reported and published, this will be done in a way that does not identify the institute, you, or your students. I will also offer you a copy of any report or publication that I write as a result of this research. If you are interested to know about the result of the research, you can provide your email or postal address on the Consent Form. I will send you a copy of any publication that results from the study.

All participants in your class who agree to take part in the study by allowing me to have access to their test results and grade information will be offered the chance to enter a draw to win a book voucher to the value of $50 NZD. If they are willing to enter the draw, they need to note this on their Consent Forms and provide their phone calls. At the end of the research, the researcher will write the names on pieces of papers in privacy and for each class, one name will be drawn randomly as the winner. The winners will be informed by the researcher through phone call. You will also receive a $50 NZD book voucher in recognition of the time you give to this study.

If you agree to participate, please fill out the Consent Form and return it to the researcher. You can discuss any issue with the researcher at any time by email or phone call or you are welcome to contact the following at the University of Auckland:

Researcher	Supervisor	Head of School of Cultures, Languages and Linguistics
Mandana Hazrat Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 3, Room 313 18 SYMONDS ST AUCKLAND 1010 New Zealand Iranian Phone number: +09399665001 Email: mhaz902@aucklanduni.ac.nz	Prof. John Read Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 2, Room 206 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 7673 Email: Ja.read@auckland.ac.nz	Professor Gary Barkhuizen Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 3, Room 318 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 8197 Email: g.barkhuizen@auckland.ac.nz

For any concerns regarding ethical issues you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Research Office, Private Bag 92019, Auckland 1142. Telephone 09 373 7599 ext. 83711. Email: ro-ethics@auckland.ac.nz

Approved by the University of Auckland Human Participants Ethics Committee on 11/06/2018 for three years, Reference Number 021365

Appendix 10: A Sample of Head of Language Institutes' Participation Information Sheet for the Pilot Study

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

Participant Information Sheet

To: Head of Language Institute

Project title: The Impact of Task Type on Language Learning

Researcher: Mandana Hazrat
Supervisor: Dr. Rosemary Erlam

I am a doctoral candidate in Applied Language Studies at the University of Auckland in New Zealand. I am interested in finding out what type of classroom activities best help students learn English. I am inviting you to allow (one of) your teachers and students to participate in a research project that will help me find this out.

I have chosen students at your institute because they are at an intermediate level of proficiency and are familiar with completing pair activities.

As part of this project, the students will receive two lessons designed for learners at an intermediate level. The lessons will be used as additional materials that are usually provided in each semester in your language institute. Before the lessons, the students will first take two tests in one session that will take about 30 minutes. In the second session, they will receive Lesson 1 that will take about one and a half hours. In the third session, the students will complete three tests that will take about an hour. The procedures of the second and third sessions will be repeated for Lesson 2 except that the students will also complete a questionnaire about the lessons in the last session that may take about 15 minutes. Therefore, the whole study will take about five hours and 45 minutes over five sessions.

I would like to ask teachers of intermediate classes to allow me to conduct the project in their classes. I will first ask them to check that the lessons are suitable for their students. If they agree and are willing to participate in this project, they will teach two lessons and administer the tests and the questionnaire. I will provide all the teaching and testing materials. Before each lesson, I will let teachers know how to carry out the procedures and I will attend the teaching and testing sessions. I also would like to access the learners' previous grades in placement tests or end-of-semester tests to have an overview of the participants' language level.

Participation in this project will be voluntary and the students and teachers can withdraw before 15th December 2017. I seek your assurance that participation or non-participation of the students and teachers will not affect their grades, employment or relationship with the school. Students who do not participate in the study will still complete all the tests and lessons as part of their instruction.

The benefit of students participating in this project includes doing individual and pair activities suitable for their level of proficiency that is likely to improve their English. Teachers' participation may also help me know how to help teachers to design effective lessons.

All information about the participants will be kept by the researcher in a locked cabinet and on a password protected computer both in Iran and in New Zealand and will be shredded or deleted after six years. If the results of this research are published, this will be done in a way that does not identify your institute, teachers, or students. I will also offer you a copy of any report or publication that I write as a result of this research. If you are interested to know about the results, you can provide your email or postal address on the Consent Form. I will send you a copy of any publication that results from the study.

All participants who agree to take part in the study by allowing me to have access to their test and questionnaire results and grade information will be offered the chance to enter a draw to win a book voucher to the value of $50 NZD. For each class, one book voucher will be given. If the participants are willing to enter the draw, they need to note this on their Consent Forms and provide their phone numbers. At the end of the research, the researcher will write the names on pieces of papers in privacy and for each class, one name will be drawn randomly as the winner. The winners will be informed by the researcher though phone call. Each teacher who participates in the project will also receive the same voucher.

If you agree to this study, please fill out the Consent Form and return it to the researcher. You can discuss any issue with the researcher at any time by email or phone call or you are welcome to contact the following at the University of Auckland:

Researcher	Supervisor and Senior Lecturer	Head of School of Cultures, Languages and Linguistics
Mandana Hazrat Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 3, Room 313 18 SYMONDS ST AUCKLAND 1010 New Zealand Iranian Phone number: +09123212684 Email: mhaz902@aucklanduni.ac.nz	Dr. Rosemary Erlam Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 2, Room 212 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 7081 Email: r.erlam@auckland.ac.nz	Professor Gary Barkhuizen Applied Language Studies and Linguistics University of Auckland ARTS 2 - Bldg 207 Level 3, Room 318 18 SYMONDS ST AUCKLAND 1010 New Zealand Tel: +64 9 923 8197 Email: g.barkhuizen@auckland.ac.nz

For any concerns regarding ethical issues you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Research Office, Private Bag 92019, Auckland 1142. Telephone: 09 373 7599 ext. 83711. Email: ro-ethics@auckland.ac.nz

Approved by the University of Auckland Human Participants Ethics Committee on 19/09/2017 for three years, Reference Number 019930

Appendix 11: A Sample of Students' Consent Form for the Pilot Study

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

ARTS

Consent Form
Student

'THIS FORM WILL BE HELD FOR A PERIOD OF SIX YEARS'

Project title: The Impact of Task Type on Language Learning
Researcher: Mandana Hazrat
Supervisor: Dr. Rosemary Erlam

I have read the Participants Information Sheet and have understood the nature of the research and why my class has been selected. I have had the opportunity to ask questions and have them answered to my satisfaction.
I agree to participate in this project and allow the researcher to use my test and questionnaire results for her research.
I understand that the researcher will obtain information about my previous performance from the language institute.
I understand that my participation in this project is voluntary, and that non-participation will not affect my relationship with teachers and the Head of the Language Institute.
I understand that I may withdraw my agreement to participate and any data traceable to me up to 15th December 2017 without giving a reason.
I understand that the test results will be kept by the researcher in a locked cabinet and a password-protected computer and will be destroyed after six years.
I further understand that my identity will be kept confidential in any publication resulting from the study.
I wish to enter the draw for the book vouchers. YES /NO
If yes, please provide your phone number: …….
I wish to receive a copy of a publication that results from the study. YES /NO

Email or postal address (if you wish to receive a copy of any report based on this study)
…….

Name:

Signature:

Date:

Approved by the University of Auckland Human Participants Ethics Committee on 19/09/2017 for three years, Reference Number 019930

Appendix 12: A Sample of Teachers' Consent Form for the Main Study, Experimental Groups

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

Consent Form
Teachers

'THIS FORM WILL BE HELD FOR A PERIOD OF SIX YEARS'

Project title: The Impact of Task Type on Language Learning

Researcher: Mandana Hazrat
Supervisor: Prof. John Read

I have read the Participant Information Sheet, have understood the nature of the research and why myself and my students have been selected. I have had the opportunity to ask questions and have them answered to my satisfaction.
I agree to participate in this project and will teach two lessons and administer tests all designed by the researcher.
I understand that the researcher will inform me about the procedures I should carry out in the classroom and will attend teaching and testing sessions.
I understand that participation in this project is voluntary and my students and I are free to withdraw before 30th of August 2018 without giving a reason and if anyone withdraw, all the information provided by her/him will be withdrawn.
I understand that my participation or non-participation will not affect my relationship with the Head of the Language Institute.
I understand that data will be kept by the researcher in a locked cabinet and a password-protected computer for 6 years, after which they will be destroyed.
I understand that my identity and my students' identities will be kept confidential in any thesis, conference presentation or journal article that is produced as a result of this project.
I wish to receive a copy of any publication that results from the study. YES/NO

Email or postal address (if you wish to receive a copy of any report based on this study)
……

Name:
Signature:
Date:

Approved by the University of Auckland Human Participants Ethics Committee on 11/06/2018 for three years, Reference Number 021365

Appendix 13: A Sample of Head of Language Institutes' Consent Form for the Pilot Study

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

ARTS

Consent Form

The Head of Language Institute

'THIS FORM WILL BE HELD FOR A PERIOD OF SIX YEARS'

Project title: The Impact of Task Type on Language Learning
Researcher: Mandana Hazrat
Supervisor: Dr. Rosemary Erlam

I have read the Participant Information Sheet, have understood the nature of the research and why my students in the language institute have been selected. I have had the opportunity to ask questions and have them answered to my satisfaction.
I give my permission for this research project to take place in my institute.
I understand that information about the participants' previous performance will be obtained from the institute.
I understand that the teachers will teach two lessons, administer tests and questionnaires all designed by the researcher. The researcher will inform the teachers about the procedures they should carry out in the classroom before each lesson and will attend teaching and testing sessions.
I understand that participation in this project is voluntary and participants are free to withdraw before 15th December 2017 and if they withdraw, all the information provided by them will be withdrawn.
I give my assurance that participation or non-participation of students and teachers in this project will not in any way affect their grades, employment or relationship with the school.
I understand that data will be kept by the researcher in a locked cabinet and a password-protected computer for 6 years, after which they will be destroyed.
I understand that the participants' identities will be kept confidential in any thesis, conference presentation, or journal article that is produced as a result of this project.
I wish to receive a copy of any publication that results from the study. YES /NO

Email or postal address (if you wish to receive a copy of any report based on this study)
……
Name:
Signature:
Date:
Approved by the University of Auckland Human Participants Ethics Committee on 19/09/2017 for three years, Reference Number 019930

Appendix 14: A Sample of Confidentiality Agreement for Raters for the Main Study

SCHOOL OF CULTURES, LANGUAGES AND LINGUISTICS
APPLIED LANGUAGE STUDIES AND LINGUISTICS
Telephone 64 9 373 7599
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

ARTS

Confidentiality agreement

Raters

'THIS FORM WILL BE HELD FOR A PERIOD OF SIX YEARS'

Project title: The Impact of Task Type on Language Learning
Researcher: Mandana Hazrat
Supervisor: Prof. John Read

I understand that information about the participants' test scores is confidential and must not be disclosed to or discussed with anyone.

Name:

Signature:

Date:

Approved by the University of Auckland Human Participants Ethics Committee on 11/06/2018 for three years, Reference Number 021365

Appendix 15: The X-Lex Test (Milton, 2009, p. 254)

Student's name…..... Classroom number ……

Do you speak any language other than Farsi and English?

Please look at these words. Some of them are real English words and some are invented but are made to look like real words. Please tick the words that you know (it is your choice to work across or down the table).

that	with	before	person	feel	round
early	table	question	effect	market	woman
darrock	believe	fine	instead	produce	group
arrive	waygood	both	century	cup	kennard
park	path	tower	gazard	wheel	whole
perform	pity	fishlock	signal	dish	earn
sweat	trick	manage	mud	cantileen	stream
pardoe	everywhere	deny	shot	gillen	independent
feeling	frequid	juice	nod	gentle	slip
diamond	press	provide	hobrow	reasonable	boil
sandy	military	candlin	staircase	daily	litholect
associate	conduct	relative	upward	publish	insult
gumm	humble	contract	mount	tube	moreover
crisis	jug	lesson	oak	alden	treadaway
limp	sumption	headlong	violent	fade	rake
trunk	mercy	anxious	horozone	arrow	feeble
sorrow	brighten	hyslop	outlet	chart	drum
difficult	manomize	antique	discuss	dam	essential
horobin	pedestrian	bullet	cliff	probable	permission
normal	cardboard	impress	refer	weather	stand

Appendix 16: Answer Key

words	Acceptable answers	words	Acceptable answers
Lesson 1		**Lesson 2**	
Agriculture	کشاورزی، زراعت	Acquire	بدست آوردن، حاصل کردن، کسب کردن
Blazing	مشتعل، شعله ور بودن، آتشی، شعله ور شدن، شعله، خیلی گرم	Addiction	اعتیاد، اعتیاد داشتن، معتاد
Climate	آب و هوا، جوی	Aim	هدف، هدف گیری
Confront	مقابله کردن، مواجهه کردن، مقابله	Beneficial	مفید، سودمند، به نفع بودن
Conserve	از صدمه محفوظ داشتن، نگه داری کردن، محافظت، نگهداری	Deficiency	کمبود، نقص، کاستی
Contamination	آلودگی، کثیفی	Devote	وقف کردن، اختصاص دادن، فدا کردن
Devastating	مخرب، ویرانگر	Diverse	متنوع، گوناگون، مختلف
Emission	نشر، بیرون دادن، خروج، دفع، پخش کردن	Enforce	مجبور کردن، وادار کردن، اجبار، اجرا کردن
Endeavour	تلاش، کوشش، سعی، تلاش کردن، سعی کردن	illogical	غیر منطقی، نامعقول
Exceptional	استثنایی، مستثنی	Impact	اثر، تاثیر، تاثیر گذار بودن
Expansion	توسعه، رشد کردن	Innovative	جدید، ابداعی
Gradual	تدریجی، قدم به قدم پیشرونده، گام به گام، مرحله به مرحله، پله پله	Intimacy	صمیمیت
Numerous	فراوان، زیاد، شماره ها، تعداد	Irritating	آزاردهنده، رنج آور، رنجیده شدن
Originate	سرچشمه گرفتن، ناشی شدن، اصلی، اصل بودن، اصلیت، اصالت، خواستگاه اولیه، شروع کردن از، ریشه، پدید آمدن، منشا	Isolation	انزوا، کناره گیری، جدا سازی، دورنگه داشتن، گوشه گیری، محدود کردن، دور کردن
Persuade	تشویق کردن، ترغیب کردن، پشت کسی بودن	Majority	اکثریت، بیشتر، گروه زیادی از مردم، عموم، بیشتر، اکثر، عده ای از مردم
Phenomena	پدیده، خارق العاده، واقعه حیرت انگیز	Participate	شرکت کردن، مشارکت کردن، فعالیت کردن
Rescue	نجات دادن، نجات، رهایی	Permit	اجازه، اجازه دادن
Vital	حیاتی، واجب، اساسی، مهم	Precious	ارزشمند، گرانبها، با ارزش
		Restrict	محدود کردن

Appendix 17: Lessons Used in the Main Study

Lesson 1: Extreme Weather

Section 1

Part A

What is the weather like in your country?

Please work in pairs and write two words that best explain the weather in each season in your country.

Spring	Summer	Fall	Winter

Part B

Which of the following conditions have happened in your country?

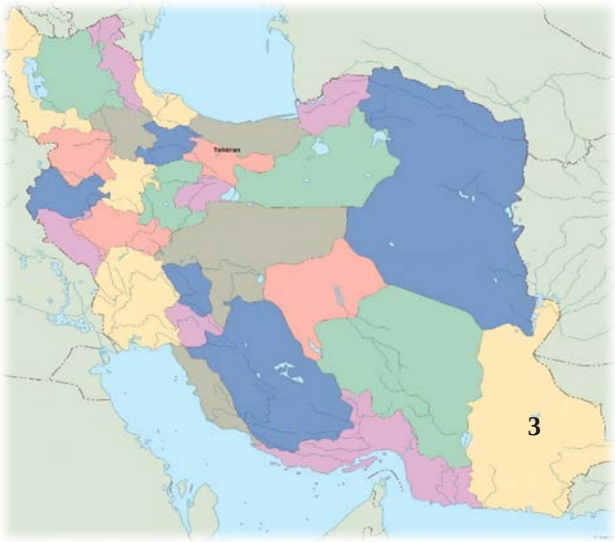
Please work individually. One of you circle the conditions that have occurred and the other circle the conditions that have not occurred. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.

For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

1- Devastating storm	2- Exceptional hot weather	3- Strong winds	4- Fall in sea levels
5- Gradual changes in weather	6- Lack of water	7- Poor agriculture	

Now, please work in pairs and look at each other's answers. Does your friend agree with you?

Please work together and write the number of each condition on the map to show where the condition has occurred (An example is number '3' on the map).



Devastating:	Causing a lot of harm	Agriculture:	Farming
Exceptional:	Very unusual	Gradual:	Not sudden

Part C

Please read the text individually and express your ideas by choosing yes, no, or not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

Now, which one do you agree with, A or B? Please work in pairs and compare your answers. If your friend does not agree with you go around the class and find someone who agrees with you. Then, put the eight reasons into three groups in the below table by writing the number of each reason in one of the rows:

Very important cause of changes in weather	
Important cause of changes in weather	
Not an important cause of changes in weather	

Originate: To happen or appear

Climate: The general weather condition in a particular area

Emission: The act of sending out of light, heat, gas …

Part D

Now, you know your friend's ideas about weather change. Please one of you read the following sentences and express your ideas by choosing yes, no, not sure. The other one of you should guess his/her friend's ideas by choosing Agree, Disagree, or Not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

What can we do to stop changes in weather?			
1-It is vital for people to use less gas and oil.	Agree	Disagree	Not sure
2-There is very little we can do to stop weather phenomena.	Agree	Disagree	Not sure
3-We can only reduce the effects of extreme weather.	Agree	Disagree	Not sure
4- The world must come together to confront changes in weather.	Agree	Disagree	Not sure
5- We should persuade people to change their habits.	Agree	Disagree	Not sure
7- We can conserve energy to stop heating of the Earth.	Agree	Disagree	Not sure

Now, please work in pairs and show your answers to each other. How many right guesses you have made?

Section 2

Please read the following table individually and tick the factors that according to the paragraph cause heating of the Earth in the table. You can look up the words you would like to know their meanings in a dictionary, but you should circle the words you look up.

Factors that cause heating of the Earth			
The sun	✓ ✗	Using energy	✓ ✗
Pollution	✓ ✗	Cloud cover	✓ ✗
Cutting trees	✓ ✗		

'Changes in weather are natural'

The Earth is affected by many natural events and one of them is weather change. Changes in weather have happened several times in the past and the Earth has often been warmer than now. Heating of the Earth is not caused by just one factor. One important factor during the past hundred years has been the Sun. The sun warms the Earth and a warmer planet will have more extreme weather such as storms and unusual rain. The earth's weather can also change because of cloud cover. Therefore, it is not true to say that human activity such as cutting trees, using too much energy, polluting the environment, and population growth cause heating of the Earth. The reason for changes in weather is mostly natural. Then, there is very little we can do for heating of the Earth. We may control numerous factors such as population expansion and contamination, but we cannot control the blazing sun. The best thing we can do is to endeavour to reduce the effects of weather events. An example is warning people before storms that can rescue people's lives and helping victims after that.

Vital:	Necessary	Persuade:	To make somebody decide to do something
Phenomena:	Events that are not fully understood	Conserve:	To use a little of something and keep it for a long time
Confront:	to deal with a problem or a difficult situation		

Section 3

Receptive Retrieval

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about changes in weather that may be different from your idea(s) about it. The words are provided at the bottom of the page. You can use each word only once.

1- The Earth's …….. *is changing and human activity is the main reason for it.*

2- Extreme weather …….. *can endanger people's lives.*

3- *Changes in weather are ……..., so people may not become aware of them in their daily lives.*

4- *Sometimes, weather events are …….. and it costs a lot to repair the effects.*

5- One reason for …….. *hot weather in some countries is air pollution.*

6- *The …….. of CO2 can happen naturally in the environment and human activity is not the only reason for pollution.*

7- *Warmer weather can make it difficult to grow plants so it can have harmful effects on ……. .*

8- The rains and storms we are experiencing may …….. *from air pollution.*

9- *To stop heating of the Earth, it is …….. for humans to change their habits.*

10- *It is necessary for all people around the words to …….. heating of the Earth.*

11- *It is not easy to …….. people to plant trees or to use less water.*

12- *By riding bikes and using less oil, people can …….. energy and stop heating of the Earth.*

O*riginate, Climate, Conserve,* D*evastating, Emission, Vital,* P*henomena, Exceptional, Confront, Agriculture, Persuade, Gradual*

Productive Retrieval

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about changes in weather that may be different from your idea(s) about it. The first letter of each word is provided for you. You can use each word only once.

1- The Earth's *climate is changing and human activity is the main reason for it.*

2- Extreme weather p*henomena can endanger people's lives.*

3- *Changes in weather are gradual, so people may not become aware of them in their daily lives.*

4- *Sometimes, weather events are* d*evastating and it costs a lot to repair the effects.*

5- One reason for *exceptional hot weather in some countries is air pollution.*

6- *The emission of CO2 can happen naturally in the environment and human activity is not the only reason for pollution.*

7- *Warmer weather can make it difficult to grow plants so it can have harmful effects on agriculture.*

8- The rains and storms we are experiencing may o*riginate from air pollution.*

9- *To stop heating of the Earth, it is vital for humans to change their habits.*

10- *It is necessary for all people around the words to confront heating of the Earth.*

11- *It is not easy to persuade people to plant trees or to use less water.*

12- *By riding bikes and using less oil, people can conserve energy and stop heating of the Earth.*

Sentence Writing

Please provide an example sentence for each of the below words.

Climate
Persuade
Exceptional
Agriculture
Confront
P*henomena*
O*riginate*
Conserve
Gradual
D*evastating*
Vital
Emission

Composition Writing

Please write a short paragraph to say whether or not you agree with the fact that heating of the Earth is an important matter for the world.

You may find the following words helpful for writing the paragraph. Each word should be used only once:

Climate, Phenomena, Gradual, Exceptional, Devastating, Vital, Emission, Originate, Agriculture, Confront, Conserve, Persuade

Lesson 2: The Internet

Section 1

Part A

What effect does the internet have on your life?

Please work in pairs and write three effects that using the internet has had on your life.

1-
2-
3-

Part B

Please read the following sentences individually and write the number of the sentences you agree with on a separate paper (for example: I agree with sentence number 1, 2, and 7). Please do not show your answers to each other. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.

For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

1-The majority of people use the internet for everyday life.
2-Young people devote much time to the internet.
3-Children give their precious time to play online games.
4-The internet provides diverse kinds of fun for children.
5-The internet is an innovative system for learning.

Now, please work in pairs. Without showing your answers, each one tells the number of the sentences she/he agrees with. The other student writes the numbers and guess his/her friends' view about the internet and tick in the below table. At the end, show what you ticked in that table to your friend. How many right guesses have you made?

My friend has a positive view about the internet	
My friend has a negative view about the internet	
My friend has neither a positive nor a negative view about the internet	

Majority:	Most of the people or things in a group	Diverse:	Very different from each other
Precious:	Valuable	Innovative:	New
Devote:	To use all or most of your time and attention to do something		

Part C

Please read the text individually and express your ideas by choosing yes, no, or not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

'A'

The internet is beneficial for children

'B'

The internet is harmful for children

The internet helps:				The internet causes:			
1-Social development	Yes	No	Not sure	5-Isolation from society	Yes	No	Not sure
2-Emotional development	Yes	No	Not sure	6-Neck and head pain	Yes	No	Not sure
3-Language development	Yes	No	Not sure	7-Addiction to computer games	Yes	No	Not sure
4-Physical development such as working with fingers	Yes	No	Not sure	8-Deficiency of physical exercise	Yes	No	Not sure

Now, which one do you agree with, A or B? Please work in pairs and compare your answers. If your partner does not agree with you go around the class and find someone who agrees with you. Then, put the eight effects into the three groups in the below table by writing the number of each effect in one of the rows

This usually happens when children use the internet	
This may or may not happen when children use the internet	
This never happens when children use the internet	

Beneficial:	Having a helpful or useful effect	Addiction:	Being unable to stop doing something as a habit
Isolation:	When a person or thing is alone and separate from others	Deficiency:	Lack of something

Part D

Now, you know your friend's ideas about the internet. Please one of you read the following sentences and express your ideas by choosing yes, no, not sure. The other one of you should guess his/her friend's ideas by choosing Agree, Disagree, or Not sure. For the search groups: Looking up the meanings of the words at the bottom of the page may help you complete the task.
For the other groups: Looking at the meanings of the words at the bottom of the page may help you complete the task.

What should and should not be done?			
1-Parents should restrict children's access to the internet.	Agree	Disagree	Not sure
2- Parents should permit children to use the internet only for doing homework.	Agree	Disagree	Not sure
3-Parents should enforce their rules for children's access to the Internet.	Agree	Disagree	Not sure
4- Parents should do nothing to stop children from using the internet.	Agree	Disagree	Not sure
5-Teachers should use the internet for its great impact on learning at schools.	Agree	Disagree	Not sure

Now, please work in pairs and show your answers to each other. How many right guesses you have made?

Section 2

Please read the following table individually and circle the hobbies as either new or old according to the information in the paragraph. You can look up the words you would like to know their meanings in a dictionary, but you should circle the words you look up.

Hobbies			
Games	New/old hobbies	Watching movies	New/old hobbies
Exercise	New/old hobbies	Family visiting	New/old hobbies
Reading	New/old hobbies	Playing with friends	New/old hobbies

'Using the internet is harmful'

Not a long time ago, life was different. Young people had enough time for family gatherings, getting some exercise, or enjoying reading a book. Since the internet has been introduced, children experienced forms of entertainment such as games, movies, and online programs. They may still enjoy old habits, but some of the new habits have caused extreme changes in their lives. Parents think that children's illogical use of the internet is irritating. Children give much of their time to use the internet with no specific aim. They rarely have time to play with friends and may never experience intimacy. This may limit their chance to have healthy relationships. Children may also spend their sleep and meal times on online activities that are not useful at all. They do not go out to participate in play and this lack of physical activity can be very harmful. To reduce the effects of the internet on children's life and health, in all families, there should be some rules to stop internet overuse. Parents should also help children to acquire healthy habits for computer use and to enjoy habit of old days.

Restrict:	To limit something	Permit:	To allow somebody do something
Enforce:	To make sure people do what they have to do	Impact:	The effect that something has

Section 3

Receptive Retrieval

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about using the internet that may be different from your idea(s) about it. The words are provided at the bottom of the page. You can use each word only once.

1- People cannot trust the internet because the …….. of online information and news are false.

2- If children …….. time to the internet they will not have time to study and play with friends.

3- It is possible for people to control the …….. of the internet on their lives.

4- …….. from the society happens when people spend their time on the internet and do not meet family and friends.

5- Like any type of …….., when people use the internet too much, they cannot stop it and do their normal life activities.

6- People who use the internet for long hours may be in danger of a …….. of physical activity.

7- If online programs stop people from doing daily activities, they have to …….. the time of using the internet.

8- In the modern world, it is not possible for parents to …….. rules for not using the internet.

9- Parents and schools should …….. children to use the internet freely.

10- Having rules for using the internet can make it …….. to learning.

11- Some schools have special programs for children with …….. learning problems.

12- Online learning and group work are not …….. ideas anymore because these days, many schools use them.

13- Children should be taught that time is …….. and it should not be spent on computer games.

Devote, Beneficial, Impact, Isolation, Restrict, Precious, Enforce, Permit, Diverse, Deficiency, Innovative, Addiction, Majority

Productive Retrieval

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about using the internet that may be different from your idea(s) about it. The first letter of each word is provided for you. You can use each word only once.

1- People cannot trust the internet because the majority of online information and news are false.

2- If children devote time to the internet they will not have time to study and play with friends.

3- It is possible for people to control the impact of the internet on their lives.

4- Isolation from the society happens when people spend their time on the internet and do not meet family and friends.

5- Like any type of addiction, when people use the internet too much, they cannot stop it and do their normal life activities.

6- People who use the internet for long hours may be in danger of a deficiency of physical activity.

7- If online programs stop people from doing daily activities, they have to restrict the time of using the internet.

8- In the modern world, it is not possible for parents to enforce rules for not using the internet.

9- Parents and schools should permit children to use the internet freely.

10- Having rules for using the internet can make it beneficial to learning.

11- Some schools have special programs for children with diverse learning problems.

12- Online learning and group work are not innovative ideas anymore because these days, many schools use them.

13- Children should be taught that time is precious and it should not be spent on computer games.

Sentence Writing

Please provide an example sentence for each of the below words.

Devote
Beneficial
Impact
Isolation
Restrict
Enforce
Permit
Diverse
Deficiency
Innovative
Majority
Addiction
Precious

Composition Writing

Please write a short paragraph to express your idea about the effects of using the internet on people's lives.

You may find the following words helpful for writing the paragraph. Each word should be used only once:

Majority, Innovative, Diverse, Precious, Beneficial, Devote, Isolation, Restrict, Impact, Addiction, Deficiency, Enforce, Permit

Appendix 18: Answer keys for Section 3

Lesson 1

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about changes in weather that may be different from your idea(s) about it.

For productive retrieval group: the first letter of each word is provided for you. You can use each word only once.

For receptive retrieval group: The words are provided at the bottom of the page. You can use each word only once.

Revised sentences with blanks	List of target words that can match the context	Acceptable answers
1- The Earth's climate is changing and human activity is the main reason for it.	Climate	Climate
2- Extreme weather phenomena can endanger people's lives.	Phenomena	Phenomena
3- Changes in weather are gradual, so people may not become aware of them in their daily lives.	Gradual, phenomena	Gradual
4- Sometimes, weather events are devastating and it costs a lot to repair the effects.	Devastating, Exceptional	Devastating, Exceptional
5- One reason for exceptional hot weather in some countries is air pollution.	Exceptional, Devastating	Exceptional, Devastating
6- The emission of CO2 can happen naturally in the environment and human activity is not the only reason for pollution.	Emission	Emission
7- Warmer weather can make it difficult to grow plants so it can have harmful effects on agriculture.	agriculture	agriculture
8- The rains and storms we are experiencing may originate from air pollution.	originate	originate
9- To stop heating of the Earth, it is vital for humans to change their habits.	vital	vital
10- It is necessary for all people around the words to confront heating of the Earth.	confront	confront
11- It is not easy to persuade people to plant trees or to use less water.	persuade	persuade
12- By riding bikes and using less oil, people can conserve energy and stop heating of the Earth.	conserve	conserve

Lesson 2

Please work individually and fill in the gaps with suitable words. The sentences express a variety of ideas about changes in weather/using the internet that may be different from your idea(s) about it.

For productive retrieval group: the first letter of each word is provided for you. You can use each word only once.

For receptive retrieval group: The words are provided at the bottom of the page. You can use each word only once.

Revised sentences with blanks	List of target words that may match the context	Acceptable answers
1- People cannot trust the internet because the majority of online information and news are false.	Majority	Majority
2- If children devote time to the internet they will not have time to study and play with friends.	Devote	Devote
3-It is possible for people to control the impact of the internet on their lives.	Impact	Impact
4- Isolation from the society happens when people spend their time on the internet and do not meet family and friends.	Isolation	Isolation
5- Like any type of addiction, when people use the internet too much, they cannot stop it and do their normal life activities.	Addiction	Addiction
6- People who use the internet for long hours may be in danger of a deficiency of physical activity.	Deficiency	Deficiency
7- If online programs stop people from doing daily activities, they have to restrict the time of using the internet.	Restrict	Restrict
8- In the modern world, it is not possible for parents to enforce rules for not using the internet.	Enforce, Restrict	Enforce
9- Parents and schools should permit children to use the internet freely.	Permit	Permit
10- Having rules for using the internet can make it beneficial to learning.	Beneficial	Beneficial
11- Some schools have special programs for children with diverse learning problems.	Diverse	Diverse
12- Online learning and group work are not innovative ideas anymore because these days, many schools use them.	Innovative	Innovative
13- Children should be taught that time is precious and it should not be spent on computer games.	Precious, beneficial	Precious

Appendix 19: Productive tests used in the Main Study

Lesson 1

Student's name…..... Classroom number ……

In the following table, please write one English word for each description. The first letter of each word is provided for you.

	Explanations	Words
1	The general weather condition in a place	Climate
2	Events that are not fully understood	Phenomena
3	Farming	Agriculture
4	Making something dirty	Contamination
5	The production and sending out of gas	Emission
6	Growth	Expansion
7	Happening slowly over a long time	Gradual
8	Causing a lot of harm	Devastating
9	Very hot	Blazing
10	Many	Numerous
11	Very unusual	Exceptional
12	Very necessary	Vital
13	To happen or appear	Originate
14	To use a little of something and keep it for a long time	Conserve
15	To save somebody or something from danger	Rescue
16	To encourage somebody to do something by giving good reasons	Persuade
17	To deal with a problem or a difficult situation	Confront
18	To try very hard to do something	Endeavour

Lesson 2

Student's name…..... Classroom number ……

In the following table, please write one English word for each description. The first letter of each word is provided for you.

	Explanations	Words
1	Being alone or separate from others	Isolation
2	The largest number of people or things	Majority
3	The effect that something has	Impact
4	The reason for doing something	Aim
5	Not having enough of something	Deficiency
6	Being unable to stop doing something	Addiction
7	Having a close relationship	Intimacy
8	Very useful	Beneficial
9	Very different from each other	Diverse
10	Something that makes others angry	Irritating
11	Not reasonable	Illogical
12	New	Innovative
13	Valuable	Precious
14	To limit something	Restrict
15	To take part in an activity	Participate
16	To give time and attention to something	Devote
17	To gain something	Acquire
18	To make sure people do what they have to do	Enforce
19	To allow somebody do something	Permit

Appendix 20: Translation Test used in the Main Study

Lesson 1

Student's name…..... Classroom number ……

It the below table, please write the meaning of each word in **Farsi**. For each word provide only one meaning that you think is the best.

1	Phenomena		10	Originate	
2	Emission		11	Conserve	
3	Climate		12	Confront	
4	Agriculture		13	Persuade	
5	Gradual		14	Rescue	
6	Exceptional		15	Contamination	
7	Vital		16	Expansion	
8	Devastating		17	Endeavour	
9	Numerous		18	Blazing	

Lesson 2

Student's name…..... Classroom number ……

It the below table, please write the meaning of each word in **Farsi**. For each word provide only one meaning that you think is the best.

1	Deficiency		11	Addiction	
2	Majority		12	Isolation	
3	Innovative		13	Impact	
4	Beneficial		14	Intimacy	
5	Diverse		15	Irritating	
6	Precious		16	Participate	
7	Devote		17	Aim	
8	Restrict		18	Logical	
9	Permit		19	Acquire	
10	Enforce				

Appendix 21: Matching Test Used in the Main study

Lesson 1

Student's name…..... Classroom number ……

Please match each word on the left with only one meaning on the right.

1- Climate D
2-Phenomena E
3-Agriculture A

A. farming
B. warming
C. safe condition
D. general weather condition
E. events that are not fully understood

4- Contamination C
5- Emission E
6- Expansion B

A. Earth
B. growth
C. making something dirty
D. the production and wasting of energy
E. the production and sending out of gas

7-Gradual A
8- Devastating C
9- Blazing E

A. happening slowly over a long time
B. happening after a long time
C. causing a lot of harm
D. very angry
E. very hot

10- Numerous B
11- Exceptional D
12- Vital F

A. most
B. many
C. general
D. unusual
E. naturally
F. necessary

13- Endeavour C
14- Originate A
15- Conserve D

I. to happen or appear
J. to put something out of sight
K. to try very hard to do something
L. to use a little of something and keep it for a long time
M. to spend a long time for doing something and do it completely

16-Rescue B
17- Persuade E
18- Confront C

A. to stop internet overuse
B. to save somebody form danger
C. to deal with a problem or a difficult situation
D. to find detailed information about something
E. to make somebody do something by giving good reasons

Lesson 2

Student's name…..... Classroom number ……

Please match each word on the left with only one meaning on the right.

1-Isolation E
2- Majority F
3- Impact C
4-Aim D

G. the ability to do something
H. the value that something has
I. the effect that something has
J. the reason for doing something
K. being alone or separate from others
L. the largest number of people or things

5- Deficiency C
6-Addiction E
7-Intimacy A

A. having a close relationship
B. not having a change in life
C. not having enough of something
D. being able to do necessary things
E. being unable to stop doing something

8- Diverse C
9- Irritating D
10- Illogical B

I. not believable
J. not reasonable
K. very different from each other
L. something that makes others angry
M. something that makes people laugh

11- Innovative A
12- Beneficial C
13- Precious E

G. new
H. rapid
I. helpful
J. possible
K. valuable

14- Restrict B
15- Devote D
16- Acquire A

I. to gain something
J. to limit something
K. to question something
L. to give time and attention to something
M. to give less value and importance to something

17- Enforce E
18- Permit C
19- Participate B

G. to ask for information
H. to take part in an activity
I. to allow somebody do something
J. to make sure something is enough
K. to make sure people do what they have to do

Appendix 22: The Shapiro-Wilk Test and Skewness and Kurtosis of Groups 1 to 10's data

Significant values were mentioned by asterisks. Non-normal distributions were highlighted.

Productive Test	df	Shapiro-Wilk Sig.	Skewness/ Std. Error	Kurtosis/ Std. Error
Group 1	17	.353	-.116/.550= -.210	-.941/1.063= -.885
Group 2	17	.693	.358/.550= .650	-.231/1.063= -.217
Group 3	17	.561	.015/.550=.027	-.921/1.063= -.866
Group 4	16	.215	.766/.564=1.358	.227/1.091=.208
Group 5	17	.695	.334/.550=.607	.156/1.063= .146
Group 6	17	.296	.100/.550=.181	-.900/1.063= -.846
Group 7	18	.749	-.117/.536= -.212	-.439/1.038=-.422
Group 8	17	.048*	-.012/.550=.021	-1.524/1.063= -1.433
Group 9	16	.035*	.026/.564=.046	-1.062/1.091= -.973
Group 10	16	.000*	2.043/.564=3.622	3.898/1.091=3.572

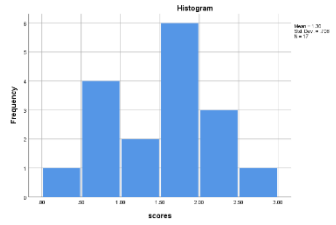
Translation Test	df	Shapiro-Wilk Sig.	Skewness/ Std. Error	Kurtosis/ Std. Error
Group 1	17	.001*	1.715/.550= 3.118	2.720/1.063= 2.558
Group 2	17	.451	-.065/.550= -.118	-.477/1.063= -.448
Group 3	17	.415	.275/.550= .5	.355/1.063= .333
Group 4	16	.149	.079/.564= .140	-1.148/1.091= -1.052
Group 5	17	.464	.296/.550= .538	1.270/1.063= 1.194
Group 6	17	.564	-.240/.550= -.436	-1.074/1.063= -1.01
Group 7	18	.034*	.968/.536= 1.805	-.039/1.038= -.037
Group 8	17	.704	.111/.550= .201	-.869/1.063= -.817
Group 9	16	.791	-.336/.564= -.595	-.604/1.091= -.553
Group 10	16	.003*	.765/.564=1.356	-.760/1.091= -.696

Matching Test	df	Shapiro-Wilk Sig.	Skewness/ Std. Error	Kurtosis/ Std. Error
Group 1	17	.116	.841/.550= 1.529	.326/1.063= .306
Group 2	17	.026*	.407/.550= .74	-.528/1.063= -.496
Group 3	17	.575	.241/.550= .438	-.717/1.063= -.674
Group 4	16	.009*	.556/.564= .985	-1.315/1.091= -1.205
Group 5	17	.118	.256/.550= .465	-.600/1.063= -.564
Group 6	17	.042*	-.135/.550= -.245	-1.641/1.063= -1.543
Group 7	18	.673	.142/.536= .264	-.414/1.038= -.398
Group 8	17	.867	.100/.550=0.181	-.142/1.063= -.133
Group 9	16	.035*	.922/.564= 1.634	.675/1.091= .618
Group 10	16	.000*	3.113/.564= 5.519	10.572/1.091= 9.690

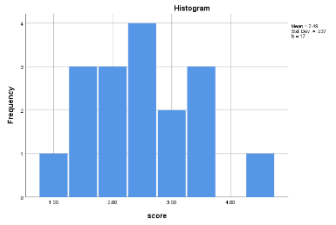
Appendix 23: Normality Histograms of Groups 1 to 10's Data

Productive Test

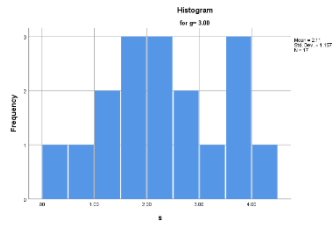
Group 1 (Approximately normal)



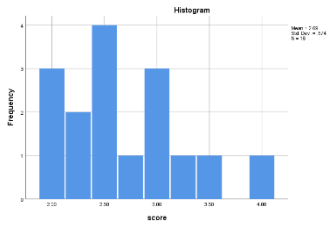
Group 2 (Approximately normal)



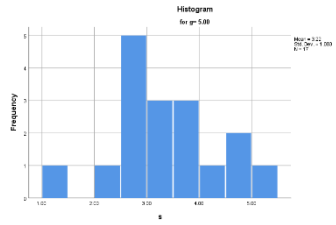
Group 3



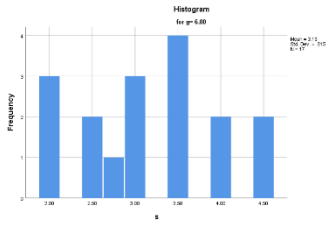
Group 4



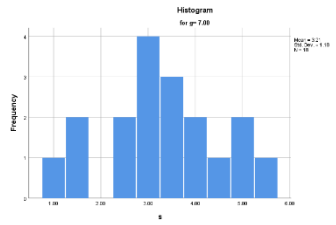
Group 5 (Approximately normal)



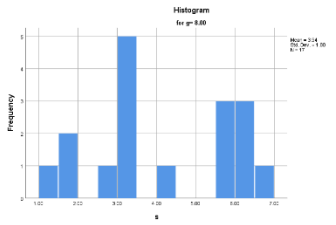
Group 6



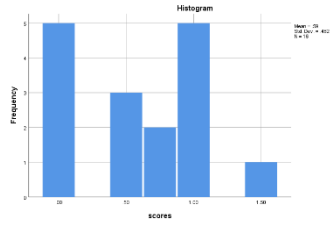
Group 7 (Approximately normal)



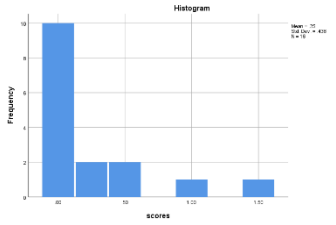
Group 8 (Approximately normal)



Group 9

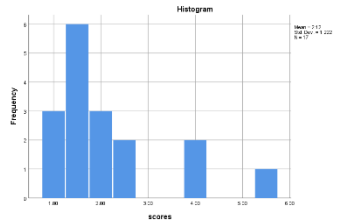


Group 10

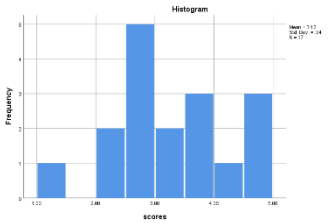


Translation Test

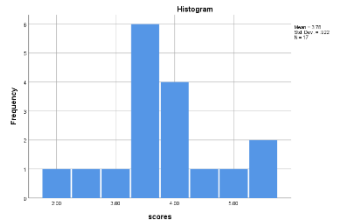
Group 1



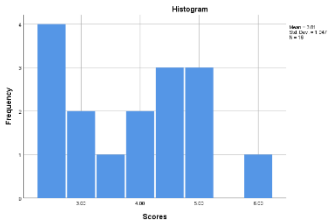
Group 2



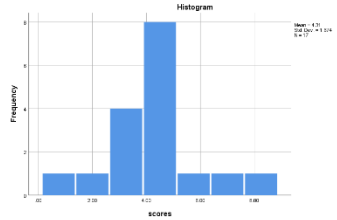
Group 3 (Approximately normal)



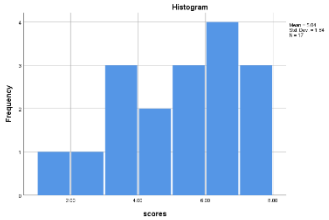
Group 4



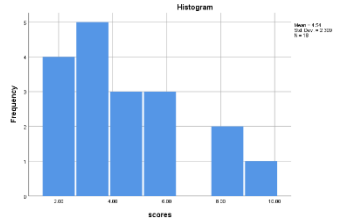
Group 5 (Approximately normal)



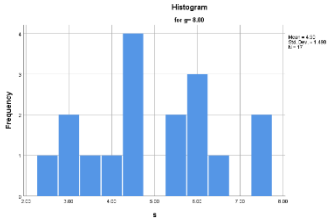
Group 6



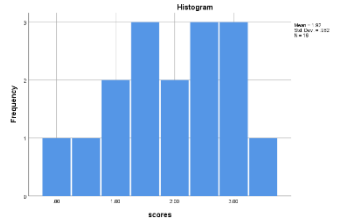
Group 7



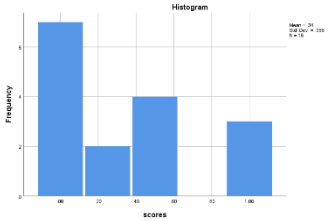
Group 8 (Approximately normal)



Group 9

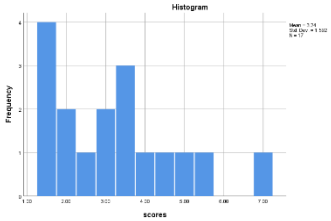


Group 10

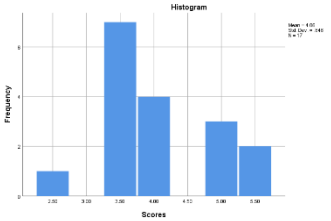


Matching Test

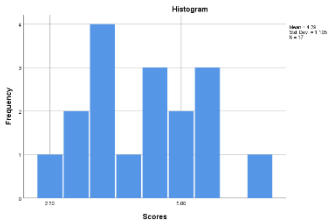
Group 1



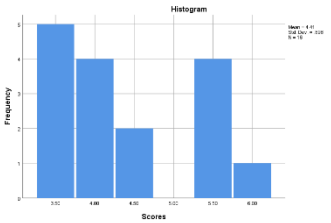
Group 2 (Approximately normal)



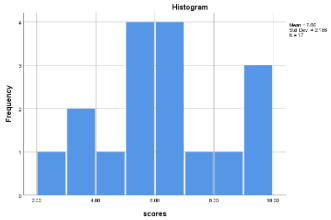
Group 3



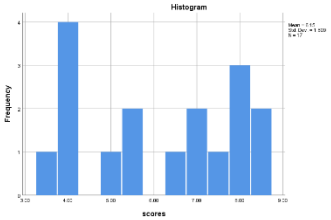
Group 4



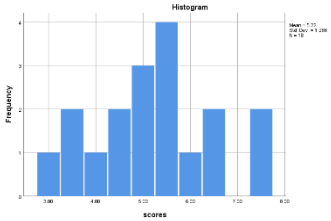
Group 5



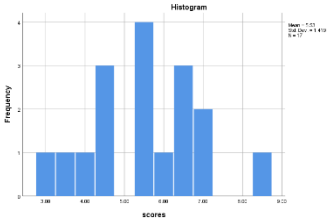
Group 6



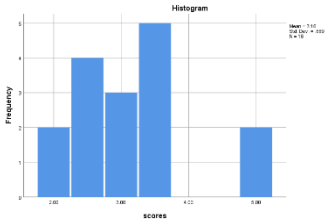
Group 7 (Approximately normal)



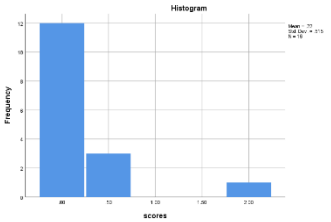
Group 8 (Approximately normal)



Group 9

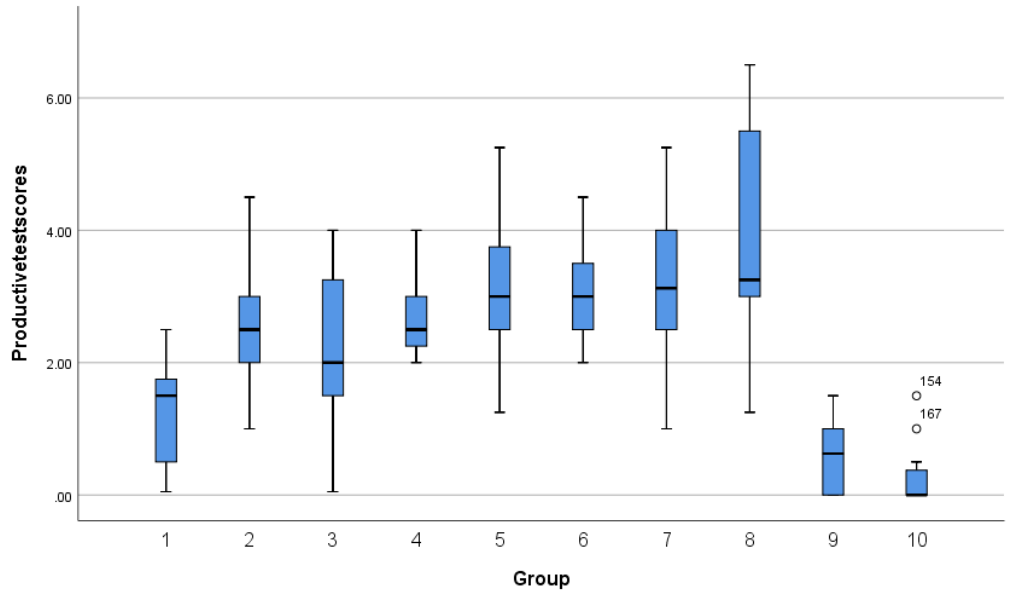


Group 10

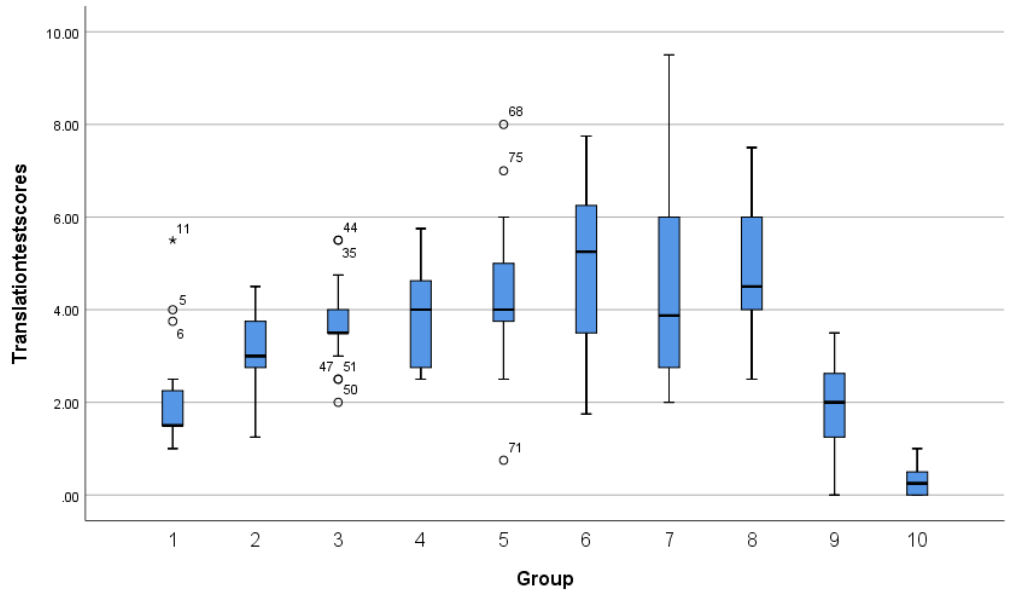


Appendix 24: Boxplots for Assessing Homogeneity of Variance of Groups 1-10's Data

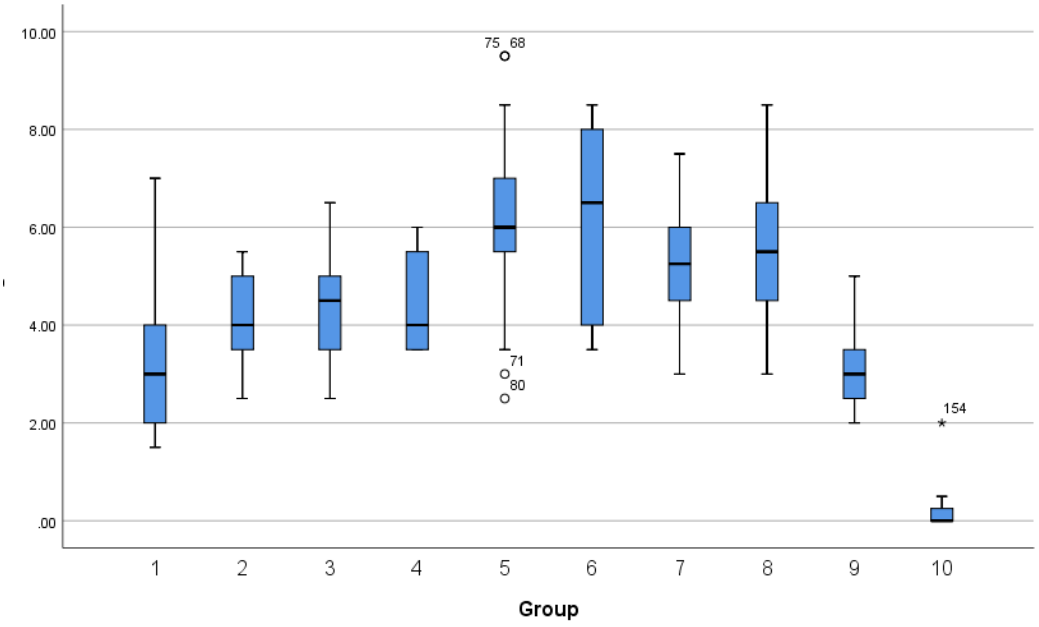
Productive test scores variance



Translation test scores variance



Matching test scores variance



Appendix 25: The Levene's Test Results

Significant values were mentioned by asterisks

		The Levene's Test		
		Levene Statistics	df	Sig.
Groups 1 and 2	Productive test scores	.847	32	.364
	Translation test scores	.038	32	.847
	Matching test scores	4.72	32	.037*
Groups 1 and 3	Productive test scores	3.47	32	.072
	Translation test scores	.078	32	.781
	Matching test scores	1.31	32	.260
Groups 1 and 5	Productive test scores	.945	32	.338
	Translation test scores	.703	32	.408
	Matching test scores	.689	32	.413
Groups 1 and 6	Productive test scores	.430	32	.517
	Translation test scores	4.232	32	.048*
	Matching test scores	1.247	32	.272
Groups 1 and 9	Productive test scores	1.320	31	.259
	Translation test scores	.000	31	.993
	Matching test scores	4.096	31	.052
Groups 2 and 3	Productive test scores	.907	32	.348
	Translation test scores	.019	32	.892
	Matching test scores	1.815	32	.187
Groups 2 and 4	Productive test scores	2.66	31	.112
	Translation test scores	.502	31	.484
	Matching test scores	.121	31	.730
Groups 2 and 5	Productive test scores	.019	32	.892
	Translation test scores	1.43	32	.239
	Matching test scores	6.34	32	.017*
Groups 2 and 6	Productive test scores	.106	32	.747
	Translation test scores	8.45	32	.007*
	Matching test scores	15.05	32	.000*
Groups 2 and 9	Productive test scores	4.142	31	.050
	Translation test scores	.101	31	.753
	Matching test scores	.002	31	.963
Groups 3 and 5	Productive test scores	.565	32	.458
	Translation test scores	1.544	32	.223
	Matching test scores	3.132	32	.086
Groups 3 and 6	Productive test scores	1.736	32	.197
	Translation test scores	8.197	32	.007*
	Matching test scores	7.018	32	.012*
Groups 5 and 6	Productive test scores	.193	32	.664
	Translation test scores	1.08	32	.307
	Matching test scores	.000	32	1.00
Groups 5 and 7	Productive test scores	.408	33	.527
	Translation test scores	1.71	33	.200
	Matching test scores	2.28	33	.140
Groups 6 and 8	Productive test scores	9.198	32	.005*
	Translation test scores	.916	32	.346
	Matching test scores	2.86	32	.100
Groups 7 and 8	Productive test scores	4.162	33	.049*
	Translation test scores	1.524	33	.226
	Matching test scores	.101	33	.753
Groups 9 and 10	Productive test scores	1.569	30	.220
	Translation test scores	10.798	30	.003*
	Matching test scores	4.916	30	.034*