Journal of English language Teaching and Learning

University of Tabriz

Volume 16, Issue 33, (Spring & Summer 2024)

On the Effectiveness of Involvement Load Components on L2 Vocabulary Learning

Zahra Memarnia

Department of English, Imam Khomeini International University, Iran
zahra.mnia98@gmail.com

Abbas-Ali Zarei (Corresponding Author)

Department of English, Imam Khomeini International University, Iran
a.zarei@hum.ikiu.ac.ir

ARTICLE INFO:

Received date:
2023.12.25
Accepted date:
2024.01.15

Print ISSN: 2251-7995
Online ISSN: 2676-6876

Keywords:

abstract words, concrete words, involvement load, vocabulary learning

Abstract

Objective: The involvement load hypothesis posits that the higher the involvement load of a task, the more effective it will be in improving students' lexical learning. It does not differentiate between the different components of involvement load (need, search, and evaluation). Nor does it assume that the type of words to be learnt has any role in the effectiveness of tasks with different involvement load indices. This study compared the effect of the components of task involvement load on the comprehension, production, and retention of concrete and abstract words.

Methods: Sixty upper-intermediate students were assigned to two groups. One group received a task in which the search component was dominant, the other group received a task (with the same overall involvement index) in which search was not present, and the evaluation component was the determining factor of task difficulty. A pretest, posttest, control group design (quasi-experimental method of research) was used to address the research questions.

Results: One-way MANOVA results on the immediate posttest were in line with ILH predictions, showing no significant differences between tasks with equal involvement indices. On the other hand, the delayed posttest results showed that in case of receptive knowledge, there was a meaningful difference between abstract and concrete vocabulary, and the search group outperformed the evaluation group. However, the results of the productive posttest showed that the evaluation group outperformed the search group in abstract words.

Conclusions: The findings can have significant implications for language learners, teachers, materials designers, and researchers.

DOI: 10.22034/elt.2024.59756.2594
Citation: Memarnia, Z; Zarei, A. A. (2024). On the Effectiveness of Involvement Load Components on L2 Vocabulary Learning. *Journal of English Language Teaching and Learning*, 16(33), 373-390. DOI: 10.22034/elt.2024.59756.2594

Introduction

Both language teachers and learners are concerned about vocabulary learning and try to discover efficient ways to develop an acceptable amount of vocabulary in language learners' long-term memory. Many researchers have tried to explain how new information is learned and retained (Webb & Nation, 2017). While L1 learners meet words in various contexts of use, and the large number of exposures facilitates words acquisition, learning a second language requires L2 learners to pay careful attention to the unknown word if they want to learn it successfully. As there is not enough context and frequency of L2 items in everyday life, they need opportunities for intentional learning (Hu & Nassaji, 2016). Moreover, it is claimed that learners' success in learning new information in second language and its retention can be partially due to the degree and depth with which the input is processed. This is what Craik and Tulving (1975) roposed as the 'depth of processing' theory, in an attempt to explain the underlying semantic and cognitive processes involved in second language learning. According to this theory, more in-depth processing of new information results in better learning and, subsequently, the connection between incoming information and pre-existing knowledge is stronger, which results in better retention of the new items (Hu & Nassaji, 2016).

Among the numerous studies that have examined the effect of this theory on learning vocabulary, the involvement load hypothesis (ILH), of Hulstijn and Laufer (2001), has attracted researchers' and teachers' attention. Laufer and Hulstijn (2001) argued that "Involvement load has a motivational – cognitive nature that predicts learners' degree of success in retention of unfamiliar words" (p.14). Based on ILH, vocabulary learning is dependent on three primary components: search, need, and evaluation. Each factor has some degree of drive.

Need refers to the requirement of word meaning for completing tasks. Need is moderate when an external authority imposes it, i.e., the instructor, or the task, like when a student has to complete a task using the words the instructor/task has asked. Need is considered strong if it is self-imposed, as a learner desires to express a concept for which s/he does not have enough appropriate forms in memory. Search is the learner's endeavor to find the meaning of a word. It is present when the learner seeks the meaning of an unfamiliar word, and absent when the learner does not have to make such an effort (e.g., reading tasks that are supplemented by glosses). Evaluation includes deciding on the word that best fits in with the sentence or context, comparing a word with other possible choices, or the special meaning of a lexical item with its different meanings (Laufer & Hulstijn, 2001). When the learner needs to recognize the differences between the words in content or the differences among various meanings of a word, evaluation is moderate. Nevertheless, if the learner has to combine novel words with pre-existing ones in a sentence, this can be a strong type of evaluation.

Although the founders of ILH believe that the involvement load is the total sum of scores attributed to each primary component, the present researchers believe that the three components may not affect vocabulary retention to the same extent. Therefore, this study attempts to study the comparative effectiveness of each component to see if the components are differentially effective on vocabulary learning. It attempts to answer these questions:

1. Is the involvement load component (search or evaluation) significantly effective on the comprehension and production of concrete vocabulary?

2. Is the involvement load component (search or evaluation) significantly effective on the comprehension and production of abstract vocabulary?

3. Is the involvement load component (search or evaluation) significantly effective on the receptive and productive retention of concrete vocabulary?

4. It the involvement load component (search or evaluation) significantly effective on the receptive and productive retention of abstract vocabulary?

Literature Review

The concept of 'task' was first developed by Prabhu (1987), who defined a task as an activity that is based on meaning. Van den Branden (2006) offered a less ambiguous definition. To him, tasks are activities in which someone gets involved to achieve a goal, for which they are required to use language. The definition provided by Ellis (2003) is more precise. He argues that a task has four features: a. it aims to involve learners in both pragmatic and semantic meanings b. it emphasizes a 'gap' in order to get students to convey information c. it requires learners to utilize their linguistic capabilities, and the most important one, d. it has a precisely defined nonlinguistic outcome.

What is somehow common in all the above conceptualizations of 'task' is that a task involves learners in some sort of mental processing. The deeper the level of the processing, the more effective the task will be. Laufer and Hulstijn (2001) proposed ILH, based on the tenets of incidental learning. ILH is defined as the motivational-cognitive construct based on which language learners' success in retaining unknown vocabulary can be predicted and estimated (Laufer & Hulstijn, 2001). According to the Dual Coding Theory, vocabulary can be encoded in memory in a couple of ways. One way is pure verbal encoding in which only the verbal information is processed. The other type supports the lexical items that have visual and sensory-motor associates or representations. Because of the high positive correlation between imageability and word's concreteness, it can be noted that abstract words-unless they are emotive- are encoded with no, or few, visual associations (Dellantonio et al., 2014). As the theory suggests, lexical items with visual associates (i.e. concrete words) can be learned easier than those that can only be verbally encoded (abstract words).

Laufer and Hulstijn (2001) also proposed two pre-assumptions for ILH. First, by keeping other factors equal, words will be remembered better if their processing requires a higher level of involvement. This implies that when other factors (like frequency of exposure) are manipulated diversely across different tasks, the findings might not be consistent with what ILH has predicted. Second, Laufer and Hulstijn (2001) hold that ILH only focuses on predicting the results of incidentally learned vocabulary. During intentional learning procedures, students may employ various techniques and strategies to boost their learning gains, and utilization of these strategies might have a noticeable effect on learning outcomes. In this way, results cannot be predicted and explained by ILH, since they might mirror the strategies that students used to

perform the given tasks rather than the cognitive processes involved in carrying out the tasks given to them.

Some studies have provided supportive evidence for ILH, suggesting that tasks with a higher level of involvement are more effective than tasks with lower involvement (Azadegan Dehkordi & Aghajanzadeh Kiasi, 2023; Eckerth & Tavakoli, 2012; Huang et. al., 2012). Nevertheless, some other studies have only partially supported the theory (Baleghizade & Abbasi, 2013; Bao, 2015; Hu & Nassaji, 2016; Kim, 2011; Yang et al., 2017). As an example, Kim (2011) conducted a study to test the claims of the ILH. The findings revealed that despite the meaningful differences among the three tasks with regard to each groups' vocabulary retention, no such difference was found in the immediate posttest. Moreover, the results supported the ILH, suggesting that both tasks with identical induced involvement were similarly effective on the advancement of not only immediate learning but also lexical retention. Mousavi et al. (2021) reported similar results about idioms learning.

Nevertheless, despite the supportive evidence mentioned above, some studies have rejected the assumptions of ILH. For instance, Hill and Laufer (2003) applied two types of tasks to test the effect of dictionary use on the retention of unknown vocabulary. The results suggested that the first task, which induced moderate evaluation, was more effective than the second task, which had the same overall involvement index but induced a strong evaluation. Such findings question the general presupposition that higher overall indices of involvement load lead to higher levels of task effectiveness.

From such controversial reports, one may gather that factors other than the sheer involvement load may be at work in determining task effectiveness. One reason for this inconsistency among the findings could be due to different weights and the impacts each factor may have in the process of vocabulary learning. Initially, the founders of ILH proposed that each factor makes an equal contribution to learning vocabulary. Nonetheless, the founders themselves, along with many other researchers, pointed out that some components might have a greater influence on vocabulary learning. For example, Laufer and Hulstijn (2001) compared the effect of three tasks with differing involvement indices on vocabulary recognition, while none of the tasks induced the search component. According to the results, the mean score for the composition writing group was the highest. However, no significant difference was found between the reading and reading plus fill-in tasks. Accordingly, they assumed that the search component might be not as effective on vocabulary learning as other components.

It is also believed that there may be variables other than the components of ILH that might account for some inconsistencies in the previous studies and influence vocabulary learning and retention. Such variables include time on task, students' vocabulary knowledge and proficiency level, and frequency of exposure, each of which are briefly reviewed below.

Some studies have yielded inconsistent results with the predictions of ILH (Laufer & Hulstijn, 2001; Martinez-Fernandez, 2008). Kim (2011) tried to explain these inconsistencies in terms of the students' proficiency levels. Although the findings of his study were compatible with the tenets of ILH, showing that proficiency level does not moderate the effect of different types of tasks, it is obvious that a more demanding task induces a higher involvement index.

Similarly, Ehsani et al. (2023) reported that the predictions of the ILH could not be confirmed in the learning of actual and pseudo words.

The term 'Time on task' is defined as the time students spend while they are engaged in activities such as sentence-writing, gap-filling, etc. (Huang et al., 2012). Some studies have shown that tasks that require more time to be completed tend to be more effective (Huang et al., 2012; Laufer & Hulstijn, 2001). As an example, Laufer and Hulstijn (2001) tested the retention (both short-term and long-term) of new words which were to be acquired incidentally by assigning the participants to three groups. Three tasks with different involvement indices and time on task were assigned to each group (reading, reading plus gap-filling, and composition-writing). The findings indicated that tasks that had higher involvement indices led to better vocabulary retention than those with lower levels of involvement. The researchers argued that the superiority of the students' performance on the composition task might not be contingent upon the time on task but to the higher induced involvement of the task, since higher involvement tasks take more time than lower involvement tasks. Laufer and Hulstijn (2001) claimed that 'time on task' should be treated as an intrinsic feature of a task. Moreover, knowing a word is not confined to the learner's knowledge about its meaning and orthographical features; it requires an understanding of numerous features such as register, collocations, deviations, grammatical functions, and constraints (Schmitt, 2008).

The term 'frequency of exposure' can be defined as the number of times leaners are either exposed to the target words or use them. As Laufer and Rozovski-Roitblat (2011) argued, acquisition of new words depends on the frequency of exposure and the quality with which these words were processed in the learners' minds. However, no specific number can be mentioned as the sufficient number of encounters necessary for vocabulary learning.

While it is generally accepted that the mentioned factors may affect task effectiveness, few, if any, studies have been done to compare the role of each of the three components of task involvement load (search, need and evaluation) in determining task effectiveness and in facilitating vocabulary learning. Furthermore, to the best of the present researchers' knowledge, the role of the type of words to be learnt in moderating such an effect has not been investigated. This study is an attempt to fill this gap.

Method

Participants

Initially, 112 female Iranian EFL students in a public high school in Qazvin, Iran took part in this study. Their age ranged from 15 to 17, and they had been learning English for 3-7 years. As the participants were students of a high school and the fact that about 85 to 90 of them had experience of attending language institute classes, their level of English proficiency could vary from lower to upper-intermediate. As a result, the researchers administered a placement test to check their level of English language proficiency. To have a homogenized sample, the researchers selected 60 students whose proficiency level was intermediate to teach the new words through online classes.

Materials and Instruments

The Macmillian Quick Placement Test was given to all the participants prior to the treatment sessions to homogenize them with regard to their language proficiency. This test contained 50 multiple choice items. Half of the items tested grammar, and half vocabulary. Its reliability, estimated using the KR-21 method, was 0.72.

The researchers began the treatment by administering a vocabulary pretest to make sure that the students had no knowledge of the selected words. The pretest consisted of 100 words (half concrete, half abstract), each contextualized in a sentence and underlined, that were taken from the books 'Oxford word skills' (Upper-intermediate), 'Vocabulary in Use' (advanced), and 'Oxford Picture Dictionary'. The students were expected to write the Persian equivalents of these words in 45 minutes. The researchers then listed those items answered by less than 5% of the students and taught those new words. For this purpose, the researchers used two different tasks.

Short-response task without marginal gloss: After learning the new words during each treatment session, students received short reading passages without marginal gloss. Then, on a separate sheet, the Persian equivalent of each new word was presented to the participants to provide their English equivalents. This task had moderate need, and search was present since they had to look up the new vocabulary in a bilingual dictionary. Evaluation was absent because the students did not need to choose between several options for each Persian equivalent. Therefore, the involvement index was 1+1+0 =2.

Fill in the blanks with marginal gloss: Here, students were provided with marginally glossed texts. After reading the texts, the students were presented with new sentences including a gap which had to be filled with one of the recently learned words. The students had to find the appropriate word to fill the gap in each sentence. The Persian equivalent of the omitted words were given as a cue near each blank space. This task induced an involvement index of 2 (+N, -S, +E), where 'evaluation' was superior to other components. Thus, the involvement index was 1+0+1=2.

Posttest: After the treatment, the students were given two tests, one for checking the comprehension and one for checking the production of the selected words. The comprehension and production posttests were administered simultaneously, immediately after the experiment. The comprehension test included 30 multiple-choice items; each item had a sentence with a blank, and the students had to choose one of the four options that was suitable for the gap. The teacher tried to choose the distractors from those incorrect meanings that students wrote for the words in the pretest to make more valid and challenging items. The production posttest included 30 gap-filling items. These items included one blank which had to be filled with one of the target words they learnt during treatment sessions. To make sure the students wrote the words that were taught (not their synonyms), an L1 equivalent was provided in parentheses near each blank. As it was mentioned earlier, each group received a different kind of task at the treatment stage. The reliability index of the vocabulary comprehension and production tests were estimated using Cronbach's alpha, and the reliability indices were 0.89 and 0.94, respectively.

The students' receptive and productive vocabulary retention was checked by giving a delayed test about a fortnight after the immediate posttest. The teacher used exactly the same items of comprehension and production. However, in order to prevent memorization effect, she re-arranged the items. The reliability of the receptive and productive retention tests was estimated using Cronbach's alpha. The results were 0.71 for receptive and 0.78 for productive retention tests.

Procedure

This study used a pretest, posttest, control group design. However, the research method was quasi-experimental because although the assignment of different groups to the treatment conditions was done quite randomly, the initial selection of the participants to take part in the study could not be done on a completely random basis. The teacher first gave a pretest to the participants to check the background vocabulary knowledge and the homogeneity of the participants. To this end, the Macmillan Quick Placement and Diagnostic Test was selected. The allocated time to this test was 35 minutes. Based on the scoring rubric, it turned out that the students' level of proficiency was mostly intermediate, though there were those with higher and lower levels as well. As a result of this homogenization process, 60 students were selected to take the treatment.

Then, the second pretest was administered. The students were asked to provide the equivalent of each word in Persian. The teacher taught the words in 10 sessions and excluded the known words from the posttests.

There was no chance for random selection of the students in in-site teaching since this study was done in a public school, and the teacher was not allowed to assign students to different classes. As a result, she decided to make online groups in a messenger application and teach the words asynchronously. The teacher assigned the students randomly to two groups. She sent two videos each week to the groups in which she taught 10-11 new words through a short text. To check if all the participants watched the video, she asked them to make a comment on the video. After 1-2 days, all students were required to complete a task. Each group received one distinct task. After watching each video, the students commented on the text of the video in 2-3 sentences. After all the students in each group commented on the video, the teacher had the participants do and send the completed task in 5-10 minutes. The teacher used two types of task that had the same overall involvement load index of 2, in order to check which component contributes more to vocabulary comprehension, production, and retention. Since there was no chance of excluding the need component from vocabulary tasks, the researchers decided to check for the difference between the evaluation and search components by holding the involvement degree of the need component constant.

One group received short-response tasks with marginal gloss in which there was a moderate need, search was present, and evaluation degree was zero since there was no need for the evaluation of new words for each blank space. Consequently, the index of involvement in this task was 1+1+0=2. The second group received a fill-in-the-blanks task with marginal gloss in which need was moderate, evaluation was moderate and search was absent. Therefore, the involvement load of this task was 1+1+0=2.

After the treatment, there were three posttests, and each test contained thirty items. One was for checking vocabulary comprehension and one for vocabulary production. In the posttests, each item was scored 1 if it was answered correctly, and 0 either if it was answered incorrectly, or it was left unanswered. Since the researchers aimed at exploring the differences between the comprehension and production of concrete and abstract words, on the immediate comprehension and production posttests, each student was given four different scores. One score for comprehension of abstract words, one for production of abstract words, one for comprehension of concrete words, and one for production of concrete words. The retention test was given a fortnight after the immediate tests to check the students' receptive and productive retention. To eliminate the memory effect, the teacher rearranged the items in both comprehension and production posttests.

After a meticulous examination of various task types and their involvement indices, the researchers had already found it impossible to design tasks with similar overall involvement indices in which only one component (need, search, and evaluation) is present. Since every vocabulary task has some degree of need component (moderate or strong); therefore, it is always present and cannot be omitted in any way. As a result, keeping the degree of the need component constant, they decided to design two different task types with an overall involvement index of 2.

Data Analysis

There were three continuous dependent variables (comprehension, production, and retention of concrete and abstract words) in each research question and two categorical independent variables, i.e., involvement load components and type of word (concrete vs. abstract). To answer each research question, a multivariate analysis of variance (MANOVA) was used.

Results

Comprehension and Production of Concrete Words

The first research question aimed to investigate the impact of involvement load components on the comprehension and production of concrete words. To this end, the students' scores on the posttests of comprehension and production of concrete words were compared. Table 1 contains the summary of the descriptive statistics.

Table 1. *Descriptive Statistics for the MANOVA on Concrete Words*

	Group	N	Mean	SD
Comp.con	Search and Need	30	9.40	2.79
	Evaluation and Need	30	8.40	3.47
Prod.con	Search and Need	30	11.10	4.07
	Evaluation and Need	30	10.13	4.00

As shown in Table 1, the evaluation group got a lower mean score than the search group on both the comprehension and production tests. To find out if the difference between the experimental groups on the immediate posttest results, one-way MANOVA procedure was

used. Before using MANOVA, the researchers checked the multivariate normality, equality of covariance, and equality of variance assumptions.

To check the multivariate normality, the maximum value for Mahalanobis distance was found to be 9.30. As we had two dependent variables, and the reported critical value for Mahalanobis distance is 13.82, there were no outliers among the scores (Tabachnick & Fidell, 2013).

To check the homogeneity of variance-covariance assumption, Box's test was used. The significance value for Box's Test was 0.24 ($P > 0.001$), meaning that this assumption was not violated. According to Table 2, Levene's test showed that no value is less than 0.05, and that we did not violate the assumption of equality of variances either.

Table 2. *Results of the Levene's Test for the Comprehension and Production of Concrete Words*

		Statistic	df1	df2	Leven Sig
Comp.con	Based on Mean	.932	1	58	.338
	Based on Median	.928	1	58	.339
Prod.con	Based on Mean	.130	1	58	.720
	Based on Median	.108	1	58	.743

Having checked the assumptions, the researchers used MANOVA. As Table 3 shows, there was no significant difference between the search and evaluation groups on the comprehension and production of concrete words [F (1, 58) =0.764, $p > 0.05$; Wilk's Lambda =0.974].

Table 3. *Multivariate Test Results for the Comprehension and Production of Concrete Words*

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Group	Pillai's Trace	.026	.764[b]	57.000	.470	.470	.026
	Wilks' Lambda	.974	.764[b]	57.000	.470	.470	.026
	Hotelling's Trace	.027	.764[b]	57.000	.470	.470	.026
	Roy's Largest Root	.027	.764[b]	57.000	.470	.470	.026

As Table 4 indicates, the obtained value for the comprehension test was non-significant [F (1, 58) =1.504, $p > 0.05$]. In addition, the outcome of the production test was insignificant [F (1, 58) = 0.857, $p > 0.05$], suggesting that there is no significant difference between these groups.

Table 4. *MANOVA Results on the Comprehension and Production of Concrete Words*

Source	Dependent Variable	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	comp.con	15.000	1.504	.225	.025
	prod.con	14.017	.857	.358	.015
Group	comp.con	15.000	1.504	.225	.025
	prod.con	14.017	.857	.358	.015

Comprehension and Production of Abstract Words

The second research question aimed to examine whether task involvement load components would have a significant effect on the participants' comprehension and production of abstract words. To check this, the researchers tabulated the comprehension and production scores on abstract words. Table 5 contains a summary of the descriptive statistics.

Table 5. *Descriptive Statistics on the Comprehension and Production of Abstract Words*

	Group	N	Mean	Std. Deviation
Comp.abs	S and N	30	10.2000	3.48791
	E and N	30	9.1333	3.78503
Prod.abs	S and N	30	8.1333	4.24860
	E and N	30	7.5667	4.02307

According to Table 5, the comprehension and production mean scores of the search group are both higher than those of the evaluation group. To see if there are significant differences among the groups, a one-way MANOVA procedure was used. Before using MANOVA, its assumptions were checked, and there was no violation. As Table 6 shows, there was no significant difference between the search and evaluation groups on the comprehension and production of abstract words [F (1, 58) = 0.638, $p > 0.05$; Wilk's Lambda = 0.978].

Table 6. *Multivariate Test Results for the Comprehension and Production of Abstract Words*

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Group	Pillai's Trace	.022	.638[b]	2.00	57.00	.53	.022
	Wilks' Lambda	.978	.638[b]	2.00	57.00	.53	.022
	Hotelling's Trace	.022	.638[b]	2.00	57.00	.53	.022
	Roy's Largest Root	.022	.638[b]	2.00	57.00	.53	.022

Moreover, Table 7 suggests that the value for the comprehension test was non-significant [F (1, 58) =1.288, $p > 0.05$]. Furthermore, the production test also yielded an insignificant result [F (1, 58) = 0.281, $p > 0.05$].

Table 7. *MANOVA Results on the Comprehension and Production of Abstract Words*

Source	Dependent Variable	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	comp.con	17.067	1.288	.261	.022
	prod.con	4.817	.281	.598	.005
Group	comp.con	17.067	1.288	.261	.022
	prod.con	4.817	.281	.598	.005

Receptive and Productive Retention of Concrete Words

The third research question aimed to examine whether task involvement load components significantly influenced the receptive and productive retention of concrete words. Table 8 presents the descriptive statistics.

Table 8. *Descriptive Statistics for the Receptive and Productive Retention of Concrete Words*

	Group	Mean	Std. Deviation	N
recretc	S and N	9.8000	3.06707	30
	E and N	8.0333	3.14570	30
Prodretc	S and N	10.6000	3.20129	30
	E and N	8.1333	2.72578	30

To see if these differences are significant, a one-way MANOVA procedure was utilized, having checked the assumptions. As Table 9 shows, a significant difference was observed between the search and evaluation groups on receptive and productive retention of concrete words [F (1,58) =5.154, $p < 0.05$; Wilk's Lambda =0.847].

Table 9. *Multivariate Test Results for the Receptive and Productive Retention of Concrete Words*

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Group	Pillai's Trace	.153	5.154[b]	2.000	57.000	.009	.153
	Wilks' Lambda	.847	5.154[b]	2.000	57.000	.009	.153
	Hotelling's Trace	.181	5.154[b]	2.000	57.000	.009	.153
	Roy's Largest Root	.181	5.154[b]	2.000	57.000	.009	.153

Table 10 shows a significant difference on both the receptive retention test [F (1, 58) = 4.851, $p < 0.05$] and the production test [F (1, 58) = 10.325, $p < 0.05$].

Table 10. *MANOVA Results on the Receptive and Productive Retention of Concrete Words*

Source	Dependent Variable	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Recretc	46.817	4.851	.032	.077
	Prodretc	91.267	10.325	.002	.151
Group	Recretc	46.817	4.851	.032	.077
	Prodretc	91.267	10.325	.002	.151

Receptive and Productive Retention of Abstract Words

To see whether the involvement load components affect the receptive and productive retention of abstract words, the researchers checked the scores on the retention test. Table 11 contains the descriptive statistics.

Table 11. *Descriptive Results on the Receptive and Productive Retention of Abstract Words*

	Group	Mean	Std. Deviation	N
recretc	S and N	7.4667	3.19194	30
	E and N	9.6000	3.13600	30
Prodretc	S and N	6.4667	3.65526	30
	E and N	8.1000	3.18780	30

To find out if the observed difference was significant, a one-way MANOVA procedure was used. Before that, all the assumptions were checked. Table 12 shows a significant

difference between the search and evaluation groups on the comprehension and production of abstract words [F (1, 58) =3.401, $p < 0.05$; Wilk's Lambda =0.893].

Table 12. *Multivariate Test Results for Receptive and Productive Retention of Abstract Words*

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Group	Pillai's Trace	.107	3.401[b]	2.000	57.000	.040	.107
	Wilks' Lambda	.893	3.401[b]	2.000	57.000	.040	.107
	Hotelling's Trace	.119	3.401[b]	2.000	57.000	.040	.107
	Roy's Largest Root	.119	3.401[b]	2.000	57.000	.040	.107

Table 13 confirms the existence of a meaningful difference between the search and evaluation groups on the receptive retention of abstract words [F (1, 58) = 6.819, $p < 0.05$]. However, despite the differential effect of the search and evaluation components on the productive retention test, the difference does not reach statistical significance level [F (1, 58) =3.402, $p > 0.05$].

Table 13. *MANOVA Results for the Receptive and Productive Retention of Abstract Words*

Source	Dependent Variable	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Recretc	68.267	6.819	.011	.105
	Prodretc	40.017	3.402	.070	.055
Group	Recretc	68.267	6.819	.011	.105
	Prodretc	40.017	3.402	.070	.055

Discussion

The results of the immediate posttests on the comprehension and production of concrete and abstract words revealed that, although the students of the search group performed better on both comprehension and production posttests, no generalizable difference was detected between the groups. This finding lends support to a number of studies (Ansarin & Kazemipour Khabbazi, 2021; Azadegan Dehkordi & Aghajanzadeh Kiasi, 2023; Bao, 2015; Laufer & Hulstijn, 2001; Mousavi et al., 2021). As Laufer and Hulstijn (2001) stated, regardless of the type of task (input- or output-based), involvement load hypothesis predicts better learning of unfamiliar vocabulary in tasks that have higher involvement indices. However, there should be no statistically significant difference between tasks with similar involvement indices. The results of this study, in the first two research questions, are indicative of the non-significant differences between tasks with similar involvement indices.

Meanwhile, Ansarin and kazemipour Khabbazi (2021) reported similar performance on the cloze deletion (+N, -S, +E) and paragraph writing (-N, -S, -E). This shows that moderate need and evaluation cannot trigger higher vocabulary development. Thus, when tasks do not induce strong evaluation or need, groups do not differ significantly on posttests. In the present study, we had two tasks, one with moderate evaluation and need, and one with search and moderate need. Similar to the mentioned study, no meaningful difference was found between our search and evaluation groups on the comprehension and production tests.

Even though the mentioned studies yielded similar results to the present study to some extent, they examined Involvement Load Hypothesis in contexts other than ours. As an example, Bao (2015) stated the effect of task type on the vocabulary comprehension and production of both male and female university students. Moreover, Ansarin and Kazemipour Khabbazi (2021) worked on multimedia contents with single and dual annotations, and Mousavi et al. (2021) studies the effect of IL on the recognition of idioms, while in this study, the researchers investigated effect of involvement load components on concrete and abstract vocabulary learning through short passages, with and without glossing, which have not been studied so far.

At the same time, compared to the outcome of the first and second research questions, a number of studies have reported contradictory results (Hu & Nassaji, 2016; Ehsani et al., 2023; Karami & Esrafili, 2021; Kim, 2011; Taheri & Rezaie Golandouz, 2021; Yanagisawa & Webb, 2021; Zou, 2017). Hu and Nassaji (2016) and Ehsani et al., 2023 examined both Technique Feature Analysis and ILH to see which is more conducive to vocabulary learning. Despite the fact that they found ILH less effective than TFA, they also reported that tasks having similar involvement load indices would not necessarily result in similar learning gains. As Hu and Nassaji (2016) claimed, one possible explanation could be different weights attributed to the components of ILH.

In another study, Tang and Treffers-Daller (2016) studied the relative effectiveness of ILH components in input-oriented and output-oriented tasks. They found significant differences among groups on the immediate posttest. In addition, their findings suggested meaningful differences between the immediate and delayed posttest scores of the groups where the dominant variable in groupings is need, or evaluation, whilst search is not a crucial component in initial vocabulary learning. The same results were obtained by Kim (2011), as she claimed much greater involvement in lexical processing is associated with strong evaluation than moderate evaluation or the other components. Contrary to these results, we found a non-significant difference between the search and evaluation groups on both comprehension and production posttests.

The contradictory results can be justified considering factors other than the involvement load index of the tasks. According to the previous studies, factors like time on task (Huang et al., 2012; Yanagisawa & Webb, 2021) and frequency of exposure (Bao, 2015) may play a role. In addition, unlike the previous researchers (Hu & Nassaji 2016; Tang & Treffers Daller, 2016), we did not have the opportunity to have male participants as well. Thus, it is possible that these findings would have been different in case we could overcome this limitation. Another reason could be the nature of our treatment. The treatment sessions had to be held asynchronously. Perhaps we could have a significant result if we held the sessions in real classrooms and had more control over participants' task completion and allocated time for each task.

The finding of the third question showed that the search group significantly outperformed the evaluation group on the posttests of productive and receptive retention. This result is in line with several studies by considering ILH stipulations (Hazrat & Read, 2021; Hill & Laufer, 2003; Yanagisawa & Webb, 2021), and it supports some others from the perspective of dictionary use and its relation with vocabulary learning (Chen, 2012; Ma & Cheon, 2018). Exploring the importance of search in comparison to other factors, Laufer and Hulstijn's (2001) findings suggest that successful vocabulary learning depends on the quality and amount of students' attention to a word. Accordingly, in reading with dictionary use, the level of

involvement is hypothetically higher; therefore, it is likely to facilitate lexical learning better than reading without dictionary use, since dictionary use includes 'need' because learners are motivated to find words in a dictionary, and 'search' for the meaning of words. It also requires moderate 'evaluation'. In addition, as Hazrat and Read (2021) maintain, 'search' is not effective by itself, though it may be so when combined with a particular evaluation type (receptive retrieval). Moreover, our findings support Ma and Cheon's (2018) findings about the positive effect of dictionary use on word learning through reading.

In the present study, the search group outperformed the evaluation group for concrete words on both receptive and productive retention tests. Unlike the previous researchers who focused mostly on the test format, task type, and frequency of exposure, we focused on the type of word. In this study, students had to do either fill-in or translation tasks after reading a short text. In the search group, the participants had to look up the words in a bilingual dictionary for doing the tasks properly. As a result, dictionary support aided learners in doing the tasks and, consequently, those learners in the search group performed better than the other group on the retention posttests. This shows the superiority of the 'search' component over the other two components in better retention of the target vocabulary, shedding light on the fact that as learners are motivated enough to look for the meaning of the word themselves instead of finding it from the text gloss, they will remember and retain it better.

Contrary to the mentioned studies, our third finding is not in line with a number of studies (Karami & Esrafili, 2021; Tang & Treffers-Daller, 2016; Yang et al., 2017). For example, Yang et al. (2017) found that sentence-writing and gap-fill groups outperformed those who completed tasks that did not induce 'evaluation'. They claimed that the numerical values given to ILH components on the delayed posttests might not affect vocabulary learning to the same extent, and 'search' is less influential than the evaluation component. However, we found the search component more effective than the evaluation component. This can be explained considering the types of word (concrete and abstract) we focused on in this study. None of the previous researchers have examined the effects of ILH components on these word types so far. Furthermore, unlike the evaluation group, students did not need to concentrate on the sentences of the tasks or the reading passage itself deeply to choose the most appropriate concrete word for each blank. With respect to concreteness effect, concrete words are learned and retained more easily than abstract one (Farely et al., 2012; Taylor et al., 2019). Thus, we can claim that since abstract words lack a sensory-imagery referent, they mostly need to be used and practiced in a rich context, so that learners can bear them more easily in their minds. Considering the relative ease of learning and retention of concrete words, and the effectiveness of looking up words in a dictionary by learners themselves, our obtained results are justifiable.

Our findings also showed a considerable difference between the evaluation and search groups on the receptive recall posttest. This finding is compatible with several studies (Laufer & Hulstijn, 2001; Yang et al., 2017; Yanagisawa & Webb, 2021). Laufer and Hulstijn's (2001) findings suggested that, although the gap-fill group (IL = 2) and the composition group (IL = 3) improved on the delayed posttest, only the latter performed better than the reading comprehension group (IL = 1). Thus, they concluded that 'evaluation' might be a key factor in vocabulary learning. In addition, Kim (2011) found that, on the delayed posttest, the comprehension group (+N, -S, ++E) performed better than the reading plus graphic organizer group (+N, -S, -E). Based on these findings, Kim held that not only do the three components of ILH not contribute to vocabulary learning to the same degree, but also 'strong evaluation'

is likely to be the most influential component. In another study by Yang et al. (2017), it turned out that although both sentence-writing and gap-fill groups outperformed the reading only and control group, no significant difference was detected between these two groups on the delayed posttest. This is probably because the learners were involved in deeper processing levels than the other tasks. Therefore, 'evaluation' is more conducive to vocabulary retention than 'need' and 'search'. Furthermore, Yanagisawa and Webb (2021) found that followed by 'need', evaluation contributed most to learning. Similar to the mentioned studies, the present researchers found that the evaluation group performed significantly better than the search group on the receptive retention test for abstract words. Regarding productive retention, the sig. value was slightly higher than the critical level on the posttest ($p > 0.05$); therefore, the difference between these groups failed to reach the significance level. In line with this, Baleghizadeh and Abbasi (2013) as well as Liu and Reynolds (2022) provided full support for the ILH.

It needs to be noted that none of the above-mentioned researchers worked on the type of word while examining the effect of ILH on word learning. Furthermore, very few of them did their experiments in a virtual environment. Based on the tenets of Dual Coding Theory about the inherent differences between the learning procedures and retention of concrete and abstract words, we can claim that since abstract words lack a sensory referent in real world, they are more difficult to learn and retain. As the evaluation group task compelled learners to analyze the sentences deeply to find the most suitable word for blanks, the learners performed better on the receptive retention posttest. However, regarding our productive retention test results, we can conclude that in the productive retention of abstract words, some other factors need to be taken into consideration. As an example, 'strong evaluation' would make the results of our productive retention test significant. Furthermore, frequency of exposure can be regarded as a determining factor in the retention test (Bao, 2015). This is especially a determining factor for abstract nouns, since learning these words is more challenging (Taylor et al., 2019). Even though the chances of word encounter in our evaluation group was higher than the search group, there is a possibility that the number of encounters for the evaluation group members were not sufficient for the learning and retention of each abstract word because the treatment sessions were held asynchronously, and the teacher could not strictly control the number of desirable exposures.

Conclusion

In this study, researchers intended to examine the ILH, supposing that the index of involvement in a task is not the single determining factor for task difficulty. However, the presence, absence, and weight of each ILH component may affect the amount of vocabulary gain drastically. One of our assumptions was that 'evaluation' is the most influential component and its presence can affect the performances of the groups, followed by 'search'. The results of our first and second research questions, however, showed no meaningful difference between the groups. This revealed that, contrary to our mentioned presupposition, ILH components do not make a difference in tasks with equal overall index, confirming ILH stipulations (Bao, 2015; Kim, 2011; Laufer & Hulstijn, 2001). The other conclusion we can draw is that only the presence of strong 'evaluation' can contribute to more effective learning.

The findings of our third research question suggest that the search group performed better than the evaluation group on the delayed tests of the recognition and production of concrete words. This outcome, which is against our immediate posttest results and ILH predictions,

reflects the importance of ILH components. We can conclude that ILH is applicable to only immediate posttest results, as a considerable difference was observed between search and evaluation groups, which shows the superiority of 'search' component over 'evaluation' on the retention of concrete words.

On the other hand, we found the evaluation component superior to the search component in the productive retention of abstract nouns. This is against our immediate posttest results, where no group was found to be superior on both comprehension and production tests. Although ILH claims that the degree of involvement load of a task is the main influential factor in vocabulary learning process, the results of our delayed posttests contradicts it. As a matter of fact, once again, we can claim that ILH can be relied on only on immediate posttests, and there is a probability of losing validity on retention tests. In addition, regarding concreteness effect, since context-availability theory, and numerous studies have found that learning and retaining concrete words are easier than abstract words, we assumed that abstract words are more context-dependent than concrete words (Taylor et al., 2019). The evaluation group's task induced higher exposure to target items compared to the task of the search group, which approves its demanding nature. Thus, we can attribute this to the superiority of evaluation group over the search group on our productive retention posttest.

It is also worthy of mention that almost all of the previously mentioned studies have examined ILH in a real face-to-face class, not in a virtual learning environment. Similar to previous researchers who found ILH predictions invalid on delayed posttests, there exist a number of factors that are fairly uncontrollable in virtual classes, which might affect students' vocabulary gain. Accordingly, researchers are uncertain about the applicability of ILH to vocabulary learning in online courses.

All in all, these findings can have important implications for stake holders. Teachers can raise students' awareness about the potential impact of input processing in deeper levels on their vocabulary gain. This implication is supported by the depth of processing theory (Craik & Tulving, 1975), which was later elaborated on by Laufer and Hulstijn (2001), who suggested the ILH. The findings of this study revealed that higher involvement index leads to greater learning gain. On the other hand, in case of concrete nouns, the search component is more influential than need and evaluation. Therefore, teachers should take this into consideration, and design tasks that induce 'search', and actively involve learners' minds.

Along with previous studies that insisted on the presence of 'evaluation', we found this component effective on the productive retention test of abstract words. Consequently, this component is recommended to be considered by language teachers and material developers.

Acknowledgments

The authors wish to thank all the students who willingly participated in the data collection process.

References

Ansarin, A. A. & Kazemipour Khabbazi, S. (2021). Task-induced involvement load and working memory: Effects on active and passive vocabulary knowledge of EFL learners in a multimedia learning environment. *Eurasian Journal of Applied Linguistics*, 7 (1), 277-302. https://doi.org/10.32601/ejal.911288

Azadegan Dehkordi Z & Aghajanzadeh Kiasi, G. (2023). Task-induced involvement loads and Iranian intermediate EFL learners' knowledge of collocations and level of motivation. *International Journal of Research in English Education*, 8(1), 29-47. http://dorl.net/dor/20.1001.1.25384015.2023.8.1.3.9

Baleghizadeh, S., & Abbasi, M. (2013). The effect of four different types of involvement indices on vocabulary learning and retention of EFL learners. *Journal of Teaching Language Skills (JTLS)*, 5 (2), 1-26. https://doi.org/10.22099/jtls.2013.1521

Bao, G. (2015). Task type effects on English as a Foreign Language learners' acquisition of receptive and productive vocabulary knowledge. *System*, 53, 84-95. http://dx.doi.org/10.1016/j.system.2015.07.006

Chen, Y. (2012). Dictionary use and vocabulary learning in the context of reading. *International Journal of Lexicography*, 25(2), 216-247. https://doi.org/10.1093/ijl/ecr031

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3), 268-294. https://doi.org/10.1037/0096-3445.104.3.268

Dellantonio, S., Mulatti, C., Pastore, L., & Job, R. (2014). Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. *Frontiers in Psychology*, 5, 708-715. https://doi.org/10.3389/fpsyg.2014.00708

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

Ehsani, M., Karami, H., & Mallahi, O. (2023). The effect of task type and word type on vocabulary learning: A comparison based on involvement load hypothesis and technique feature analysis. *Iranian Journal of Applied Language Studies*, 15(1), 169-190. https://doi.org/10.22111/IJALS.2023.45695.2355

Ellis, R. (2003). *Task-based language teaching and learning*. *RELC Journal*, 34(1), 64-81. https://doi.org/10.1177%2F003368820303400105

Hazrat, M. & Read, J. (2021). Enhancing the involvement load hypothesis as a tool for classroom vocabulary research. *TESOL Quarterly*, 56(1), 387-400. https://doi.org/10.1002/tesq.3051

Hill, M., B., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *International Journal of Applied Linguistics*, 41(2), 87-106. http://dx.doi.org/10.1515/iral.2003.007

Hu, H. 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, 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

Karami, H. & Esrafili, M. (2021). The impact of task type and involvement load index on Iranian EFL learners' incidental vocabulary learning and retention. *Journal of Language Horizons*, 5(1). 251-266. https://doi.org/10.22051/lghor.2020.31501.1311

Kim Y. J. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning,* 61(1), 100-140. https://doi.org/10.1111/j.1467-9922.2011.00644.x

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

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

Liu, S., & Reynolds, B.L. (2022). Empirical support for the involvement load hypothesis (ILH): A systematic review. *Behavioural Sciences, 12,* 1-23. https://doi.org/10.3390/ bs12100354

Ma, J. H., & Cheon, H. J. (2018). An experimental study of dictionary use on vocabulary learning and reading comprehension in different task conditions. *International Journal of Lexicography,* 31(1), 29-52. https://doi.org/10.1093/ijl/ecw037

Mousavi, M., Zarei, A. A., & Ahanghari, S. (2021). The effects of task focus and involvement load on idioms recognition. *Journal of Modern Research in English Language Studies* 8(4), 159-181. https://doi.org/10.30479/JMRELS.2021.15357.1893

Prabhu, N. S. (1987). *Second language pedagogy.* Oxford University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363. https://doi.org/10.1177/1362168808089921

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson Education.

Taheri, S., & Rezaie Golandouz, G. (2021). The effect of task type on EFL learners' acquisition and retention of vocabulary: an evaluation of the involvement load hypothesis. *Cogent Education*, 8(1), 1915226. https://doi.org/10.1080/2331186X.2021.1915226

Taylor, R.S., Francis, W.S., Borunda-Vazquez, L., & Carbajal, J. (2019). Mechanisms of word concreteness effects in explicit memory: Does context availability play a role? *Memory and Cognition,* 47(7), 169–181. https://doi.org/10.1177/2158244017730596

Tang, C., & Treffers-Daller, J. (2016). Assessing incidental vocabulary learning by Chinese EFL learners: A test of the involvement load hypothesis. In G. Yu & Y. Jin (Eds.), Assessing Chinese learners of English (pp. 121–149). Palgrave Macmillan. https://doi.org/10.1057/9781137449788_7

Van den Branden, K. (2006). *Task-based language teaching: From theory to practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667282

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Yanagisawa, A. & Webb, S. (2021). To What Extent Does the Involvement Load Hypothesis Predict Incidental L2 Vocabulary Learning? A Meta-Analysis. *Language Learning,* 71(2), 487-536. https://doi.org/10.1111/lang.12444

Yang, Y., Shintani, N., Li, S., & Zhang, Y. (2017). The effectiveness of post-reading word-focused activities and their associations with working memory. *System*, 70, 38–49. https://doi.org/10.1016/j.system.2017.09.012

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