The Impact of Task Type and Involvement Load Index on Iranian EFL Learners' Incidental Vocabulary Learning and Retention

Research Article

Hossein Karami*[1]
Mahnaz Esrafili[2]

Received: 2020-05-21 | Revised (3): 2020-08-10 | Accepted: 2020-11-10

Abstract

Vocabulary knowledge has a significant role in communication. Therefore, a plethora of research has investigated the effect of various factors on the acquisition of L2 vocabulary. This empirical study aimed to examine the predictions of Laufer and Hulstijn's (2001) involvement load hypothesis (ILH) by considering the impact of task type and involvement index on Iranian EFL learners' incidental vocabulary learning. The ILH predicts that tasks with the same involvement load should lead to equal vocabulary gains. To this end, forty-five upper-intermediate learners

[1] Assistant Professor, Department of English Language and Literature, Faculty of Foreign Languages and Literatures, University of Tehran, Tehran, Iran. (Corresponding Author); hkarami@ut.ac.ir
[2] MA graduate, Department of English Language and Literature, Faculty of Foreign Languages and Literatures, University of Tehran, Tehran, Iran; mahnazesrafili1@gmail.com

DOI: 10.22051/LGHOR.2020.31501.1311

were given one of four different tasks with the same involvement index (i.e., reading comprehension plus dictionary use, cloze-exercise plus dictionary use, inferencing, and sentence writing plus meanings of target words in L1 + L2). Participants' receptive and productive knowledge of target words were tested immediately after fulfilling the tasks and three weeks later. Mixed-plot ANOVAs were run to compare the performance of the participants on the tasks. The results revealed that, contrary to the theory's predictions, the performance of the learners on the four tasks was significantly different although the tasks had the same involvement indices. The results suggest that depth of processing and involvement are important factors in vocabulary learning. However, it also seems that the relative importance of the components of involvement might require reconsideration.

Keywords: incidental vocabulary learning, involvement load hypothesis, depth of processing, task, second language acquisition

Introduction

Learning second language (L2) vocabulary is normally considered as a critical aspect of language learning. Many researchers (e.g., Laufer, 1993; Nation, 2001; Schmitt, 2008) believe that a word is the basic unit of language learning. Although the best means of achieving effective vocabulary learning is still unknown, attempts have been made to explore ways of improving students' vocabulary knowledge. One of the most-influential hypotheses has been Laufer and Hulstijn's (2001) involvement load hypothesis (ILH).

According to Laufer and Hulstijn (2001), vocabulary learning and retention is dependent on involvement load that includes three components: need (N), search (S), and evaluation (E). The theory claims that the load of a task is the sum of the prominence degrees of these three components. Consequently, tasks with higher degree of the involvement index are more effective on vocabulary learning and retention compared to tasks with lower load (Laufer & Hulstijn, 2001).

This study attempts to test the predictions of the ILH. There were two main purposes behind the present study. The first objective was to see whether the same involvement index, regardless of the type of task, leads to equal vocabulary gains. Second, whether different task types with similar involvement indices but different distribution of components have equal impact on word learning and retention of Iranian EFL learners.

Review of Literature

Explanations have been offered in the literature as to why more engagement with new vocabulary items lead to higher vocabulary learning and retention. Craik and Lockhart (1972) offered the depth of processing hypothesis (DOP). They argued that the chances of a new word to be stored in the long-term

memory are increased by the degree of learners' attention and the amount of deep processing of that word. It means that deeper processing of a vocabulary item will lead to better learning and retention of that item.

According to the DOP theory, a deeper level of processing entails strong and detailed attention, encoding, elaboration, storing, and retrieving the information. Although almost all researchers agree with the DOP, the theory has been criticized by some researchers (e.g., Baddeley, 1978; Eysenck, 1978; Nelson, 1977). According to the critics, the problem lies in the meaning of the level of processing, and the ability to distinguish between the higher and lower levels of processing. The two major disadvantages of the DOP and the necessity of offering an operational definition of different levels of mental processing led Laufer and Hulstijn (2001) to introduce the ILH. Laufer and Hulstijn accept the importance of the level of processing. However, they assert that by operationalizing the cognitive notions they will be able to omit the weaknesses of the DOP. Based on the ILH, it is possible to predict the degree of involvement load of each task by considering three factors: need, search, and evaluation.

Need is a motivational component. "Need is concerned with the *need to achieve*" (Laufer & Hulstijn, 2001, p. 14). It is active when completion of a task is required and can come in two degrees: moderate + (1) or strong ++ (2). Need is moderate when an authority like a teacher asks students to complete a task, use a word in a sentence, or write a composition by targeted vocabularies (extrinsic motivation). On the other hand, need is considered to be strong when it is internally obligated by the learners, when they aim to do a task to acquire or to search a word in a dictionary in order to carry out a task. The search component is either present (i.e., search + [1]) or absent (i.e., search – [0]). Search can be present if learners make an effort to find the meaning of an unfamiliar L2 word by consulting a dictionary, asking teachers, or peers (negotiation), or inferring.

Evaluation is another cognitive component, and like need, has two prominence degrees. When learners need to make a decision and select the most appropriate word by comparing that word with other words (by its meaning, form, or suitability) according to the specific context, the evaluation dimension is active and its involvement load is moderate and is denoted by "evaluation +". On the other hand, if after choosing the suitable word among other options, it is required to use that target word to combine it with additional words to generate an original sentence or context, evaluation is strong, and is symbolized by "evaluation ++".

Laufer and Hulstijn (2001) compared three tasks: reading (N+, S-, E-), fill-in-the-blanks (N+, S-, E+), and composition writing (N+, S-, E++), to examine the impact of the task type on the vocabulary learning and retention. Their findings showed that composition writing with the involvement index of three outscored the other two tasks. Besides, the fill-in-the-blanks task group outperformed the reading task in incidental word learning and retention. Overall, their results strongly supported the ILH.

Keating (2008) examined beginner Spanish language learners through three tasks with varying involvement loads. Keating noted that "time-on-task is an important factor to consider when evaluating the relative effectiveness of vocabulary learning tasks" (p. 382). The results revealed that higher involvement load leads to higher vocabulary learning.

Similarly, Kim (2008) conducted a study including two experiments by considering different levels of proficiency and task types. Actually, the first experiment was a partial reproduction of Laufer and Hulstijn's (2001) experiment. Kim (2008) carried out the study by giving the participants three tasks that had different involvement indices. She also controlled time-on-task. The results showed that, compared to the gap-fill and reading groups, the composition group had the greatest immediate posttest performance. The evidence in the first experiment suggested that there is no relationship between vocabulary learning and proficiency level.

The purpose of the second experiment was to examine whether different tasks with similar involvement index affect vocabulary learning and retention of students with different proficiency levels. The main reason behind performing the second experiment was to test Laufer's (2005) assumption that writing sentences and composition writing demonstrate the same degree of involvement load. The findings of the second phase of the study showed that the two tasks with the same involvement load have comparable effects on vocabulary learning and subsequent retention, and students' proficiency levels had no influence on initial vocabulary acquisition.

Kim (2008) kept the task time fixed. That is, time-on-task was fixed for all three groups. However, in Keating's (2008) experiments, time-on-task was not fixed. According to Bao (2015), another important difference between the experiments of Kim (2008) and Keating (2008) is the type of post hoc test.

Nassaji and Hu (2012) examined the effect of task-induced involvement load on language learners' inferring word meanings from context. The learners were asked to read a text and infer the meanings of 10 unfamiliar target words. Think aloud protocols were used in order to gain information about the different strategies employed by the learners. The first text included multiple-choice glosses for the target words in the margins, with the lowest task-induced involvement load (need+, search-, and evaluation+). The second text was a regular text with the target words and with moderate involvement load (need+, search+, and evaluation+). The target words in the third text were designed in derivationally different forms from the original target words, (need+, search++, evaluation++) with the highest task-induced involvement load. The results supported the hypothesis that reading texts with more cognitive effort increases learners' initial learning and retention.

There are also other studies that have failed to provide supportive evidence for the ILH. Zou (2017) investigated the evaluation component of the ILH. She examined vocabulary acquisition through three tasks: cloze exercise with involvement index of 2 (i.e., need +, search -, evaluation +), and sentence writing

and composition writing with the similar involvement index of 3 (i.e., need +, search -, evaluation ++).

Zou's (2017) findings only partly supported the ILH. In initial vocabulary learning, the participants who had completed the tasks with higher involvement load outperformed the participants who had completed the task with lower involvement load. However, tasks with similar involvement load did not lead to the same amount of vocabulary learning on the posttest. The result was in contrast with the predictions made by the hypothesis. According to the hypothesis, since sentence writing and composition writing have the same involvement load, the performance of both groups of participants should be similar. Nevertheless, the two groups did not have the same performance on the posttest. Therefore, she proposed an augmented evaluation framework: "I propose an augmented evaluation framework, suggesting that evaluation induced by the cloze-exercises should be 'moderate +', evaluation induced by sentence-writing 'strong ++', and evaluation induced by composition-writing 'very strong +++" (p. 18).

In another experimental study, Laufer (2003) gave her participants three different tasks that had similar involvement loads: reading comprehension and dictionary consultation to find the meaning of unfamiliar words (need+, search+, evaluation+, involvement index = 3), writing original sentences with the target words (need+, search-, evaluation++, involvement index = 3), and completing sentences after finding up the meaning of unfamiliar words (need+, search+, evaluation+, involvement index = 3). All three tasks had the same involvement index, and she expected to find equal vocabulary gains. Nevertheless, the findings of the experiment revealed that, on the immediate posttest, reading comprehension group received the lowest scores. Yet, the difference between sentence writing and sentence completion groups was not statistically significant. Likewise, the sentence completion group outperformed the other groups on the delayed posttest. Therefore, the evidence found by Laufer's (2003) experiment was not in keeping with the predictions made based on the ILH.

Although a few studies (e.g., Keating, 2008; Kim, 2008; Nassaji & Hu, 2012; Zou, 2017) have been done to reexamine the ILH and the impact of different types of tasks on incidental word learning, researchers have mainly focused on the comparison between the effectiveness of different types of tasks with varying involvement loads. However, few studies have directly examined the effect of different task types with similar involvement load on incidental vocabulary learning. Hence, this study was conducted to inspect the degree to which the assumptions underlying the ILH are plausible and the extent to which teachers and language learners can rely on the predictions made by the hypothesis. In addition, by considering Laufer and Hulstijn's (2001) theory and their proposal that "the impact of incidental learning of *search* might be lower than that of *need* and *evaluation*" (p. 21), another question emerged. Whether a task that consists of need+, search-, and evaluation+ could provide the same amounts of vocabulary learning compared to a task that consists of need+, search+, and

evaluation-. Laufer and Hulstijn (2001) state that more empirical studies in diverse educational settings, proficiency levels, and task types are required for examining the predictions made by the ILH.

In light of previous studies on the effectiveness of different types of tasks on incidental vocabulary learning, this study compares the influence of four different task types, inducing similar level of involvement load on incidental vocabulary learning. Thus, it seeks answers to the following questions:

1. Do tasks with the same involvement load but a different distribution of the components involved lead to equal vocabulary gains in Iranian EFL learners?
2. Do different tasks that involve the search factor have the same effect on initial vocabulary learning and retention of Iranian EFL learners?
3. Is there any statistically significant difference between input-oriented and output-oriented tasks in learning and retention of new vocabulary items by Iranian EFL learners?

Method

The Tasks

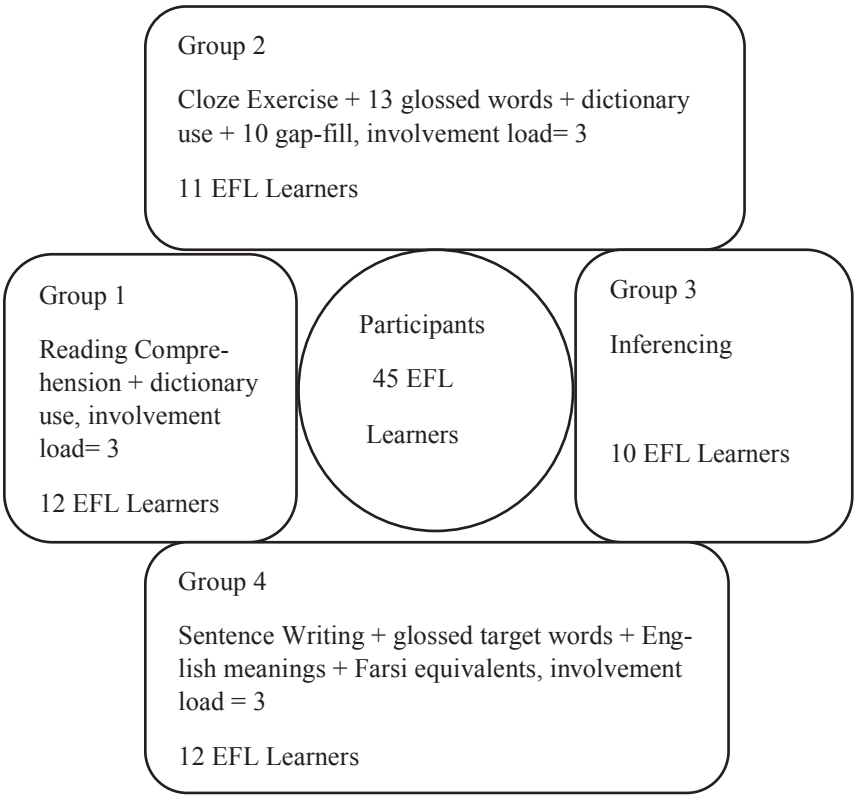
In keeping with the objectives of the study, forty-five upper-intermediate English as a foreign language (EFL) learners were randomly assigned to four groups to complete one of the four different task types with similar involvement index (i.e., 3): reading comprehension + dictionary use (N+, S+, E+), cloze-exercise + dictionary use (N+, S+, E+), inferencing (N+, S+, E+) with no dictionary use, and sentence writing + meanings of target words in L1 + L2 (N+, S-, E++). Intact classes were used in this study. Hence, the selection of the participants was done through convenience sampling. The assignment of the groups to each of the treatments, however, was randomly done. In order to assess their vocabulary learning and retention, the participants were provided with an immediate posttest following completing the task and a delayed posttest three weeks later. The posttest was designed by the researchers and was the same test used as the delayed posttest. The reliability of the scores obtained from the posttest was estimated through Cronbach Alpha and was 0.87. Cronbach alpha was selected as it is one of the most frequently used reliability indices.

Participants

First, a total number of 53 (35 females and 18 males) EFL language learners were selected though convenience sampling. However, the participants were not aware of the real nature of the study and they were not informed about an upcoming delayed posttest. Hence, eight of them did not show up on the day the delayed posttest was given. Thus, the final number of participants in the current study was 45 Iranian EFL language learners that were from 4 intact classes from the same institute. The age of participants ranged from 17 to 34.

Based on their scores on the final exam of their previous semester, they were reported to be upper-intermediate. However, in order to be more confident of the homogeneity of the participants, an Oxford Placement Test (OPT) was also administered. Descriptive statistics were obtained from the data. The results confirmed that the participants were at the upper intermediate level. We could not randomly select each individual and assign him/her to different classes. Hence, intact classes were selected, but the assignment of each class to one of the four tasks was totally random. Each class was required to complete one of the four different types of task (i.e., Reading comprehension, Cloze Exercise, Inferencing, and Sentence Writing) (as shown in Figure 1).

Figure 1
Different Groups and Their Defined Tasks



Procedure

For the purpose of the study, ten target words (apprehensive adj., oration n., vexed adj., spawn v., envision v., abate v., caveat n., assiduous adj., stymie v., and

divulge v.) were selected. In order to remove any possible influence of the grammatical category, different parts of speech (i.e., noun, verb, and adjective) consisting of 5-12 letters were employed. Since all groups received the same vocabulary items, we assumed that the use of different parts of speech could not differentially affect their performance.

One week before the study, seven learners similar to the target group were asked to complete Min's (2008) vocabulary knowledge scale to indicate the level of their familiarity with the target words (see Table 1). The objective of this test was to determine the familiarity of the participants with the target words. Since the administration of the test to the final participants could severely affect the results, we had to select a similar group of participants. The results revealed that the learners had no knowledge of the target words.

Table 1
Min's Modified Vocabulary Knowledge Scale (2008)

	[A]	[B]	[C]	[D]
Target words	I don't remember having seen this word before.	I have seen this word before, but I don't know what it means	I know this it means	I can use this word in a sentence. (write a sentence)

Following Laufer and Hulstijn (2001), four different tasks with similar involvement loads were used (see Table 2).

Table 2
Different Type of Tasks and Their Degrees of Involvement

Type of task	Groups	Need	Search	Evaluation	Involvement index
Reading comprehension no glossary + dictionary use	12	Moderate (1)	Present (1)	Moderate (1)	3
Cloze exercise no glossary + dictionary use	11	Moderate (1)	Present (1)	Moderate (1)	3
Reading comprehension by inferencing	10	Moderate (1)	Present (1)	Moderate (1)	3
Sentences Writing +marginal glosses	12	Moderate (1)	Absent (0)	Strong (2)	3

The participants were provided with two vocabulary tests: a multiple-choice recognition test to recognize word form, and Folse's (2006) modified vocabulary knowledge scale (VKS). The tests were run once immediately after task

completion and again three weeks later to measure the students' retention. At this stage, the participants met the target words once via completing their respective tasks. Hence, using Min's (2008) VKS with its first item (i.e., I don't remember having seen this word before) did not make sense. Therefore, in order to test their initial target words learning and retention, instead of Min's (2008) four-item VKS, Folse's (2006) VKS with three items was adopted (see Table 3).

Table 3
Folse's (2006) Modified Vocabulary Knowledge Scale

	[A]	[B]	[C]
Target words	I cannot remember the meaning of the word	I can remember the meaning of the word it means	I can write a sentence by using this word, for example:

The scoring system for the multiple-choice recognition test is simple. Each incorrect target word form received a score of zero and the correct form was given a full score (i.e., 1). The scoring system for Folse's (2006) VKS was used based on some previous studies (e.g., Keating, 2008; Laufer & Hulstijn, 2001; Paribakht & Wesche, 1993; Zou, 2017). A meaning received a score of zero if it was completely incorrect, a half score if a plausible semantic equivalent of the target word was provided, and a full score if the provided meaning was correct. As for the sentences, the criteria used in the studies of Laufer and Hulstijn (2001), Paribakht and Wesche (1993), and Zou (2017) were employed. A score of 0 was given to a sentence if the target word appeared in a completely incorrect sentence (grammatically and semantically). A half score was given to a sentence if the target word was used semantically appropriately but ungrammatically, and a full score of 1 was given to a sentence if the target word was used in a semantically and grammatically appropriate context. As there were ten target words, in the multiple-choice recognition test, full marks equaled 10, and in the VKS test, full marks equaled 20 (10 for meanings and 10 for sentences).

One of the researchers and another experienced teacher separately scored the two posttests. Inter-rater reliability estimates for the immediate and delayed posttests (as assessed through Pearson correlation) were 97.7% and 96.9%, respectively.

Results

The descriptive statistics for the immediate and delayed post-test are given in Table 4. It is clear that the participants showed evidence of impressive vocabulary learning and retention in all four tasks. It appears that the largest immediate recognition scores of the participants were in the cloze task. On the other hand, sentence writing had the lowest scores. As for immediate production, sentence writing had by far the highest scores while reading had the lowest scores. In the delayed recognition test, the cloze test led to the highest scores

while the sentence writing had the lowest scores. This is exactly what also appeared in the immediate phase. The order of the tasks from the largest scores to lowest scores both in immediate and delayed recognition aspects is the cloze, reading, inferencing, and sentence writing.

As for the immediate production tasks, the participants got the highest and lowest scores in sentence writing and reading, respectively. On the other hand, and surprisingly, inferencing had the highest scores in the delayed production phase. In addition, reading had the lowest scores.

Table 4
Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Sentence Writing	Immediate Recognition	12	5.00	10.00	7.000	1.348
	Immediate Production	12	.00	15.25	6.229	5.138
	Delayed Recognition	12	2.00	10.00	6.250	2.598
	Delayed Production	12	.00	3.75	.708	1.300
Inferencing	Immediate Recognition	10	6.00	10.00	7.500	1.640
	Immediate Production	10	.00	5.50	2.725	1.902
	Delayed Recognition	10	4.00	10.00	6.800	2.044
	Delayed Production	10	.00	3.00	1.250	1.178
Cloze	Immediate Recognition	11	6.00	10.00	8.818	1.168
	Immediate Production	11	.00	3.50	2.068	1.225
	Delayed Recognition	11	5.00	10.00	8.454	1.863
	Delayed Production	11	.00	1.00	.295	.458
Reading	Immediate Recognition	12	7.00	9.00	8.000	.738
	Immediate Production	12	.00	3.25	1.458	1.152
	Delayed Recognition	12	3.00	10.00	7.167	2.368
	Delayed Production	12	.00	1.00	.146	.310

In order to see the statistical significance of the difference, a mixed ANOVA was run. The grouping of the participants (based on the tasks they worked on) was considered as the between-subject factor while the immediate and delayed post-tests for both recognition and production were taken as within-subject factors.

Before running the mixed ANOVA, the assumptions were checked. The Box's M test of the homogeneity of covariance matrices was significant at $p < 0.001$. Hence, the assumption was rejected. In such situations, Pillai's Trace should be interpreted rather than the usual Wilks' Lambda (see Pallant, 2016). Before interpreting the results, it must be checked that the interaction terms are not significant. A significant interaction term means the main effects cannot be interpreted as they appear. It is clear that all interaction terms are significant (see Table 5). Hence, the results of the mixed ANOVA cannot be interpreted and separate analyses must be done.

Table 5
Mixed ANOVA Results

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Recognition	Wilks' Lambda	.080	474.391	1	41	.000	.920
Recognition * Group	Wilks' Lambda	.490	14.243	3	41	.000	.510
Production	Wilks' Lambda	.511	39.247	1	41	.000	.489
Production * Group	Wilks' Lambda	.760	4.312	3	41	.010	.240
Recognition * Production	Wilks' Lambda	.774	11.952	1	41	.001	.226
Recognition * Production * Group	Wilks' Lambda	.793	3.568	3	41	.022	.207

The next step in the analysis was to run four one-way between-subjects ANOVAs. The results are displayed in Table 6. It is clear that except for delayed recognition, all other differences are significant. Note that the number of participants in this study was small due to the difficulties of data collection. This certainly affects the results of significance testing. Specifically, significance testing is very sensitive to sample size. Thus, with small sample sizes, moderate or even large effects might not be significant. In order to overcome this problem, effect size tests are usually reported. This comes in the last column of Table 6. In order to interpret the Eta squared values, Cohen's (1988) guidelines are usually considered. Based on his rule of thumb, 0.01 is considered as a small effect, 0.06 as a moderate effect, and 0.14 as a large effect. Here, the Eta squared values indicate that all effects show a large effect except for delayed recognition which shows a moderate effect.

Table 6
Results of One-Way ANOVAs

		Sum of Squares	df	Mean Square	F	Sig.	Eta squared
Immediate Recognition	Between Groups	20.441	3	6.814	4.356	.009	0.24
	Within Groups	64.136	41	1.564			
	Total	84.578	44				
Immediate Production	Between Groups	162.790	3	54.263	6.310	.001	0.32
	Within Groups	352.604	41	8.600			
	Total	515.394	44				
Delayed Recognition	Between Groups	29.667	3	9.889	1.947	.137	0.12
	Within Groups	208.244	41	5.079			
	Total	237.911	44				
Delayed Production	Between Groups	7.839	3	2.613	3.127	.036	0.19
	Within Groups	34.264	41	.836			
	Total	42.103	44				

A significant F-test would mean that at least one of the comparisons is significant. In order to see which of the comparisons are significant, the Sheffe post-hoc test was run. The results of the post-hoc analyses are displayed in Table 7. Note that the results for delayed recognition are not displayed because the main effect was not significant, though it showed a moderate effect. Based on

the post-hoc analyses, it is clear that in the immediate recognition phase, the difference between sentence writing and cloze is the only significant difference. In the immediate production, on the other hand, the differences between sentence writing and the three other tasks is significant. The differences between reading, cloze, and inferencing are not significant. Finally, in the delayed production, the difference between inferencing and reading is the only significant difference.

Table 7
Post-Hoc Analyses

Dependent Variable	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.
Immediate Recognition	Sentence Writing	Inferencing	-.500	.535	.832
		Cloze	-1.818*	.522	.013
		Reading	-1.000	.510	.294
	Inferencing	Cloze	-1.318	.546	.138
		Reading	-.500	.535	.832
	Cloze	Reading	.818	.522	.491
Immediate Production	Sentence Writing	Inferencing	3.504	1.255	.065
		Cloze	4.161*	1.224	.016
		Reading	4.771*	1.197	.004
	Inferencing	Cloze	.657	1.281	.966
		Reading	1.267	1.256	.797
	Cloze	Reading	.610	1.224	.969
Delayed Production	Sentence Writing	Inferencing	-.542	.391	.595
		Cloze	.413	.382	.761
		Reading	.562	.373	.525
	Inferencing	Cloze	.954	.399	.144
		Reading	1.104	.391	.061
	Cloze	Reading	.150	.382	.984

Discussion

According to Laufer and Hulstijn (2001), the superiority of a task in vocabulary acquisition "is determined by the involvement load it induces," (p. 21) (i.e., a task with involvement index of 3 will be more effective than a task with involvement index of 2). Our findings show that each of the tasks used in this study led to different levels of learning and retention although they had the same involvement index. Thus, our findings are in contrast to the predictions of the ILH and previous research findings which argued that the involvement index alone determines the degree of task effectiveness in vocabulary learning. Laufer and Hulstijn (2001) clearly state that other factors such as task types (e.g., input-oriented versus output-oriented task) have no significant influence on the effectiveness of a task in vocabulary learning and retention. The results of our study do not confirm their claim.

Each of the four tasks showed differences in both immediate and delayed posttests. The effectiveness of all four tasks showed significant differences even though they had the same involvement load. Hence, the results do not support the ILH. Moreover, the cloze-exercise task had the best results and the sentence

writing task had the worst performance on the immediate recognition posttest (i.e., 8.81, and 7). The difference between the cloze-exercise task and the sentence writing task was the only significant difference in immediate recognition phase. The cloze task is followed by reading comprehension, and inferencing tasks.

In the immediate productive knowledge test, the sentence writing group, which had strong evaluation++, performed significantly better than all three other groups which had moderate evaluation+. These results are in keeping with Laufer and Hulstijn's (2001) experiment concerning Dutch-English learners. In addition, the results support the findings of Kim's (2008) and Zou's (2017) studies. In Kim's study, the sentence writing and composition writing groups (both with strong evaluation) performed significantly higher than the other two groups (i.e., gap-fill group with moderate evaluation and reading group with no evaluation).

In the delayed recognition test, the participants in cloze test gained the highest scores while the participants in sentence writing got the lowest scores. This is exactly what also took place in the immediate recognition test. Therefore, the order of the tasks from the highest scores to lowest scores both in immediate and delayed recognition tests was as follows: the cloze-exercise, reading comprehension, inferencing, and sentence writing.

As for the retention of productive word knowledge, the participants obtained the highest and lowest scores in inferencing and reading comprehension, respectively (i.e., 1.25, and .14). While there were differences between all four tasks in delayed productive vocabulary knowledge, only the difference between inferencing and reading comprehension tasks was significant.

The results of the study demonstrate that the difference between cloze and sentence writing was the only significant difference in immediate recognition test. This may be attributed to the lack of the cognitive component of search in sentence writing task. Since the component of search is absent in sentence writing, such processing does not take place and, consequently, the participants in this group obtained the lowest score in both immediate and delayed recognition tests. Besides, the finding of the current study revealed that, although the component of search may affect immediate and delayed recognition of words' form, different realizations of search have similar effects on both incidental word learning and retention (i.e., consulting a dictionary, asking a teacher or peer, or inferring).

In addition, the findings in the immediate productive test showed that the strong type of the evaluation component induces much greater level of processing than the other components (i.e., moderate and strong need, search, and moderate evaluation) at least at the first stages of word learning. It means that strong evaluation ++ possibly increases learner's initial vocabulary learning (Kim, 2008; Zou, 2017). Although there are differences between all four tasks in the delayed recognition test, none of them is significant. Note that, due to the difficulties of data collection, the number of participants in this study was small

and significance testing was very sensitive to sample size. Hence, with small sample size, moderate or even large, effects might not be significant. Nevertheless, this finding is in line with the ILH in that different types of task that induce similar involvement index generate similar amount of retention.

Furthermore, in the delayed productive phase, the participants in inferencing group outscored the other three groups and the difference between inferencing and reading comprehension tasks was significant. It may be the case that inferencing has a facilitative impact on vocabulary learning (Nassaji & Hu, 2012; Nation, 2001; Webb & Chang, 2015). It seems that inferring the meaning of unknown words induces much greater level of processing that leads to less memory loss and the deeper learning and retention of word knowledge.

One possible explanation for the lowest mean scores of the reading comprehension task may be ventured here. Reading a text and looking up its unknown words to answer some reading questions lead to less retention of productive word knowledge possibly because they do not deeply involve the learner. This might be true if we accept the ILH hypothesis. Remember that in the present study, even the sentence writing group, which obtained the highest productive vocabulary knowledge, experienced the greatest drop after the three-week interval. Therefore, multiple exposures to new words is required in order to improve retention of incidentally learnt vocabulary. Moreover, compared to three other tasks, the inferencing task suffered the slightest decline in productive vocabulary knowledge from the immediate to delayed posttests. This is evident from the changes in means scores from the immediate to delayed posttests. As explained before, the nature of the inferencing is making connections between learners' background knowledge and available clues in text. It seems that a higher degree of involvement takes place through processing the new words. This may be the main reason behind less memory loss over time in the inferencing task.

Conclusion

Based on the results of the study, although Laufer and Hulstijn (2001) assume an equal impact for all three components of the hypothesis, it appears that equal importance should not be attached to these components. In addition, the findings reveal that different realizations of search induce similar amount of word learning and retention in both immediate and delayed posttests. It also seems that being an output- or input-oriented task cannot guarantee the level of processing and vocabulary learning. It appears that the involvement index alone cannot predict the efficacy of vocabulary tasks. Therefore, the operational definitions of the components of involvement regardless of other factors that make a task more or less effective might require reconsideration.

Our findings also show that, due to the lack of frequent exposures and repetition of the new words, all groups suffered a memory loss from immediate posttest to delayed posttest. Thus, in an EFL setting where language learners suffer from the lack of exposure outside of English classes, designing tasks that

create an opportunity to expose students to the new lexical items could be effective (Nation & Wang, 1999). Therefore, by considering multiple exposures as a key factor in changing temporary vocabulary knowledge to permanent knowledge, teachers could play an influential role in this process

The progress in vocabulary learning may take place via doing tasks with strong evaluation, writing-focused tasks, or even deliberate repetition of new target words after completing tasks. Some previous studies (e.g., Folse, 2005; Keating, 2008) pointed out the positive influence of repetition via instruction. It is suggested that teachers design a variety of writing-based tasks that have strong contextual clues for inferring meaning of unfamiliar words that can encourage learners to produce sentences, paragraphs, or compositions with target words to help them to better anchor the words in memory.

It is clear that incidental and intentional approaches to vocabulary learning have to be considered as equal partners that need each other (Schmitt, 2008). Due to the restriction of time and the difficulty to teach vocabulary, teachers can teach target words through deliberate, long-term, and effective incidental word learning programs. Furthermore, a wide range of exposure to newly learned words and intentional approaches is required to consolidate and expand incidentally learnt lexical items.

References

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik & Lockhart's framework for memory research. *Psychological Review, 85*(3), 139-152.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

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

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology, 69*(2), 157-169. https://doi/10.1111/j.2044-8295.1978.tb01643.x.

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

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%2F1362168808089922.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(1), 285-325. https://doi.org/10.1111/j.1467-9922.2011.00644.x.

Laufer, B. (1993). The effect of definitions and examples on the use and comprehension of new words. *Cahiers de Lexicologie*, *63*(2), 131-142.

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. In S. H. Foster-Cohen, M. Garcia-Mayo, & J. Cenoz (Eds.), *Eurosla Yearbook* (Vol. 5, pp. 223-50). John Benjamins.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 16*(3), 307-322. https://doi.org/10.1093/applin/22.1.1.

Min, H. T. (2008). EFL vocabulary acquisition and retention: Reading plus vocabulary enhancement activities and narrow reading. *Language Learning, 58*(1), 73-115. https://doi.org/10.1111/j.1467-9922.2007.00435.x.

Nassaji, H., & Hu, M. (2012). The relationship between task-induced involvement load and learning new words from context. *IRAL, 50*(1), 69-86. https://doi.org/10.1515/iral-2012-0003.

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

Nation, I. S. P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2), 355-380.

Nelson, T. O. (1997). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior, 16*(2), 151-171. https://doi.org/10.1016/S0022-5371(77)80044-3.

Pallant, J. (2016). *SPSS survival manual* (6th ed.). Open University Press.

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

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363. https://doi.org/10.1177%2F1362168808089921.

Webb, S., & Chang, C. S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition, 37*(4), 651-675. https://doi.org/10.1017/S0272263114000606.

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%2F1362168816652418.