The effect of task type on EFL learners’ acquisition and retention of vocabulary: an evaluation of the involvement load hypothesis

Sahar Taheri¹ and Ghafour Rezaie Golandouz²

Abstract: This study set out to inspect Laufer and Hulstijn’s (2001) motivational-cognitive construct of task-induced involvement. Involvement load hypothesis predicts that tasks with the same involvement loads lead to the same amount of learning. To this end, three output task types which induce equal involvement indices of 3 were employed to examine the extent to which they promoted learners’ vocabulary acquisition and retention. Participants were 78 Iranian EFL learners at the intermediate level from three intact classes who were assigned to one of the three experimental groups. Participants in the first group were required to make sentences using the target words; and participants in the second group were required to summarize the texts using the target words; while participants in the third group were required to predict the end of the two incomplete stories using the target words. The results revealed that the three tasks significantly affected vocabulary learning. However, it was found that the three tasks with the same loads of involvement led to significant differences in the learning and retention of the target words. It was concluded that when task-effectiveness is concerned, involvement load is not the only determining factor.

Subjects: Theories of Learning; Classroom Practice; Language Teaching & Learning

ABOUT THE AUTHOR
Sahar Taheri obtained her MA degree in TELF from Islamic Azad University, Science and Research Branch in Tehran, Iran. She has nine years of experience in teaching general English and has been an IELTS instructor since 2019. Her main area of research includes task-based language teaching, reading, and vocabulary acquisition, as well as Form-focused Instructions. Dr. Ghafour Rezaie Golandouz is an assistant professor of Teaching English as a Foreign Language (TEFL) at the Islamic Azad University, Garmser Branch, Garmser, Iran. His main research interests include cognitive aspects of second language acquisition, research in applied linguistics, and language assessment. He has published, reviewed, and edited numerous research articles.

PUBLIC INTEREST STATEMENT
Clearly, vocabulary is considered to be the most important element in language learning. Therefore, many hypotheses have so far been put forward to identify the factors which are important in vocabulary learning. One of these hypotheses is Involvement Load Hypothesis (ILH) which indicates that the higher the level of involvement in the learning tasks, the higher the amount of vocabulary learning. This study was done to test this hypothesis. To this aim, three tasks which had similar ILH indices were chosen and investigated with 78 Iranian EFL learners. The results showed that all the three tasks significantly affected vocabulary learning. However, it was found that the three tasks with the same loads of involvement led to significant differences in the learning and retention of the target words. It was thus concluded that ILH can be partially true when it comes to vocabulary learning and retention.
Keywords: Task; task types; involvement load hypothesis; vocabulary learning; vocabulary retention

1. Introduction

Vocabulary knowledge is an essential component of second language learning. Knowledge of vocabulary in a language entails knowledge of its spoken and written forms, meaning, and the ability to use it in a context (Nation ISP, 2001). A review of the related literature indicates that the field is replete with investigations concerning vocabulary learning, with many attempting to identify the most effective tasks in this area (e.g., Hulstijn et al., 1996; Newton, 2001; Wesche & Paribakht, 2000). The Involvement Load Hypothesis (ILH) is among the most commonly accepted views in determining task effectiveness in the realm of vocabulary learning. As ILH has a highly predictive power concerning the evaluation of the effectiveness of tasks in vocabulary learning, the ILH promises great pedagogical implications. Based on this hypothesis, the effectiveness of a given task in learning unfamiliar words is related to the degree of its involvement, which is the sum of the prominence degrees of three components: need, search, and evaluation. Since the proposal of this hypothesis, numerous studies have examined it and its predictive power. Many researchers (e.g., Amini & Maftoon, 2017; Hulstijn & Laufer, 2001; Namaziandost et al., 2020; Pourakbari & Biria, 2015) have found supportive evidence for ILH, while some researchers have found contradictory evidence for this hypothesis (e.g., Folse, 2006; Laufer, 2003). Therefore, due to previous contradictory research findings, further investigations are required to shed more light on the predictive power of ILH in the realm of vocabulary acquisition.

2. Vocabulary learning, input, output, and ILH

2.1. Input, reading, and vocabulary learning

One of the most controversial issues in L2 vocabulary learning is the extent to which learners’ exposure to written input can yield effective vocabulary learning. The importance of comprehensible input in vocabulary learning was first put forward by Krashen’s (1989) Input Hypothesis, which suggests that like L1 vocabulary learning, the bulk of L2 vocabulary is learnt through exposure to comprehensible input. However, the results of various studies (e.g., Horst et al., 1998; Nation ISP, 2001; Schmitt, 2000) indicate that vocabulary gain through reading is small and a process prone to errors. Peters et al. (2009) pointed out that one of the crucial factors determining the retention of the words through reading is that the words should be processed elaborately. In a similar vein, in his Noticing Hypothesis, Schmidt (1990) claims that in order for learning to occur, learners should attend to at least some aspects of the input, and without attention, no input becomes intake. In line with these positions, the results of many studies have shown that although exposing learners to reading is necessary for vocabulary learning to occur, mere exposure to reading is not a sufficient condition for vocabulary acquisition to transpire. Schmidt (1990) maintains that reading should be accompanied by some learning activities so that noticing of some aspects of the input (e.g., grammar and/or vocabulary) occurs and consequently acquisition takes place.

2.2. Output and vocabulary learning

The importance of output production has been put forward by Swain (1993), who argued that in addition to being exposed to comprehensible input, learners should practice the new form in their own production. According to Swain (2005), in the process of producing output, learners notice their linguistic shortcomings and errors while attempting to convey meaning (comprehensible output), and they have the opportunity to form and test their hypotheses. In this regard, the findings of a number of empirical studies have proven that reading comprehension accompanied by learners’ production is more effective than reading only in vocabulary learning (e.g., Hulstijn & Laufer, 2001; Laufer, 2003; Paribakht & Wesche, 1997).
2.3. Involvement load hypothesis

The involvement load hypothesis acknowledges that as “depth of processing” (Craik & Lockhart, 1972) suggests, the extent to which a new piece of information is remembered is dependent on the quality of attention that is paid to its various aspects. The more in-depth processing a task requires for a newly learned item, the better it retains in our long-term memory, where depth here refers to semantic involvement (Craik & Lockhart, 1972). As Craik and Lockhart (1972) maintain, semantic involvement refers to the depth of processing a task imposes on learners to recognize and retain vocabulary items. The involvement load hypothesis is an attempt to operationalize the concept of depth of processing and holds that the more involved the learners are in a word learning task, the greater they will recall the new words later. Therefore, the higher the involvement index of a task, the more effective it is in vocabulary learning. According to this hypothesis, three major components construct the involvement load of a task: need, search, and evaluation. Need is a motivational, non-cognitive component and a desire to do a task. Need is moderate (need+) when learners are externally exposed by the teacher or the task to know the meaning of one word and is strong (need++) when learners are intrinsically motivated to know a word. Search is a cognitive component and is defined as “the attempt to find the meaning of an unknown L2 word or trying to find the L2 word form expressing a concept […] by consulting a dictionary or another authority” (Laufer & Hulstijn, 2001, p. 14). Search is absent (search-) when learners are provided with the meanings of the new words and is present (search+) when learners look for the L2 word form while trying to express a concept. Evaluation is the other cognitive component and it is moderate (evaluation+) when learners have to compare the meaning of one word with other words to assess whether it fits the given context (fill in blank exercises with words provided). Evaluation is strong (evaluation++) when learners need to decide how to combine the words to generate the context for the new words such as making sentences using the new words (Laufer & Hulstijn, 2001).

The involvement load of a task is the combination of need, search, and evaluation considering their “degrees of prominence” (Hulstijn & Laufer, 2001, p. 544). According to Hulstijn and Laufer (2001), the degree of prominence refers to how moderate and strong the elements of need, search, and evaluation are based on the task type. For instance, As Hulstijn and Laufer (2001) hold, need is moderate when it is imposed by an external agent. For example, when learners are asked to use a word in a sentence by the teacher, the degree of prominence is moderate as the need is external. Based on ILH, tasks with higher loads are more effective in vocabulary learning. However, one of the limitations of the ILH is related to the evaluation component of this hypothesis which does not clearly explain why only two degrees of prominence are accorded to different methods of evaluation (Zou, 2017). For instance, sentence-making and composition-writing which have been frequently examined in various studies are both allotted strong evaluation since the learners are supposed to generate the context of the words. However, some researchers (Kim, 2008; Zou, 2017) have noted that composition-writing is a more effective activity in vocabulary learning, and involves deeper processing of the target words compared with sentence-making.

2.4. Research on the involvement load hypothesis

Since the proposal of the involvement load hypothesis, various empirical studies have been conducted to test its accuracy and predictive power. Many studies have found supportive evidence for ILH. For instance, Hulstijn and Laufer (2001) found that comprehension task plus filling in the target words (need+, search-, evaluation+) was less effective than composition writing task (need+, search-, evaluation++), but more effective than reading comprehension with marginal glosses (need+, search-, evaluation-). Hulstijn and Laufer (2001) argue that outperformance of the composition task group can provide support for Swain’s Output Hypothesis as this task made the learners stretch their linguistic resources. They further state that the Involvement Load Hypothesis does not predict that all output tasks will culminate in better results compared to input tasks. “It predicts that higher involvement in a word induced by the task will result in better retention, regardless of whether it is an input or an output task” (p. 552). To provide evidence for Involvement Load Hypothesis, Laufer and Hulstijn (2001) did an analysis of several investigations
which compared task effect on learning and concluded that the more effective tasks had a higher involvement load than the less effective tasks.

Keating (2008) investigated the effects of three tasks on vocabulary retention. Keating (2008) used three tasks including reading comprehension with marginal glasses, reading comprehension plus fill-in, and sentence-making. The results substantiated the predictions of the Involvement Load Hypothesis as retention was highest in the sentence writing task, lower in the reading plus fill-in task, and lowest in the reading comprehension task. Keating (2008) highlights the discrepancy between the finding of his study and that of Hulstijn and Laufer (2001) and argues that the better performance of the participants in Hulstijn and Laufer (2001) study is that in their task learners were required to employ the target words to write an original composition whereas in his study participants used them to write original sentences which were unrelated. Keating (2008) further argues that “it might be the case that producing connected discourse involves more elaborate processing of the target words than producing disconnected sentences” (p. 739).

Kim (2008) conducted a study composed of two experiments in which the participants were supposed to complete two tasks including composition writing and sentence making with the same involvement loads of 3. The study also sought to probe the effectiveness of three vocabulary tasks with different levels of task-induced involvement. The findings of this study suggested that both tasks —composition writing and sentence making— were equally beneficial in vocabulary acquisition. The results also revealed that a higher level of learner involvement during the task promoted more effective initial vocabulary learning and better retention of the new words. Based on the results, Kim (2008) suggested that the extent to which varying degrees of each individual component of task load (need, search, and evaluation) is conducive to an overall involvement load might not be the same. Kim (2008) concludes that teachers should consider involvement load in vocabulary task design as tasks with a higher involvement are more beneficial for retention of new words compared to tasks with a lower involvement. In line with Kim (2008), Nassaji and Hu (2012) observed that the tasks in which the participants had to infer the meanings of target vocabulary items and make some derivational changes to them with the involvement index of 5 was more effective than the task with inferring the meanings and no options provided for the learners with the index of 3. The task of inferring while the options were provided for the learners with the index of 2 was the least effective. In another study, Namaziandost et al. (2020) aimed to compare the impact of high involvement load tasks versus lack of involvement load tasks on vocabulary learning. The results of statistical analyses indicated that exposing learners to high level of involvement load tasks played a significant role in developing English vocabulary. Moreover, the results showed that vocabulary retention acquired through tasks with high level of involvement load was not significant although high involvement load was more effective in vocabulary retention compared to tasks with lack of involvement load. Similarly, Pourakbari and Biria (2015) examined the efficacy of task-induced involvement in incidental lexical development. The results indicated that tasks with the highest degree of involvement load contributed most to lexical development. Likewise, Amini and Maftoon (2017) explored the effect of task involvement load on vocabulary learning. The results showed that learners benefited more from the task with higher involvement load.

Some researchers have even demonstrated the value of writing tasks with high loads of involvement. For instance, the findings of Laufer and Rozovski-Roitblat (2011) study showed that reading plus writing tasks was more effective than reading plus dictionary use. Similarly, Pichette et al. (2012) indicated that writing tasks with strong evaluation were significantly better than reading tasks with zero or moderate evaluation.

Some have found partial supports to the involvement load hypothesis. For example, in a study conducted by Zou (2017) three tasks- cloze exercises, sentence-writing, and composition-writing - with involvement indices of 2, 3, and 3 respectively were investigated. The results of this research were partially consistent with the predictions of the hypothesis: the composition-writing and sentence-making tasks with greater involvement loads of 3 led to significantly
better word learning than cloze-exercises with a lower load of 2, while composition-writing was significantly more effective than sentence-writing despite having equal involvement indices of 3. In another study, Rahmani et al. (2018) probed the effect of four task types including creative sentence writing, story writing, summary story writing, and sentence writing with different indices of task-induced involvement load on EFL learners’ recognition and recall of unfamiliar vocabulary items. The results revealed that there were significant within-group and between-group differences among the four task types both in immediate and delayed posttests. More specifically, the creative sentence writing group outperformed the other three groups, both in immediate and delayed posttests. The imaginary story writing, summary story writing and sentence writing groups were at the second, third, and fourth places, respectively. In a similar line of research, Un-udom (2018) examined the effect of initial vocabulary learning and retention through employing the involvement load hypothesis through two tasks including building a sentence with a target word shown in the marginal gloss and constructing a sentence by searching the meaning on a bilingual dictionary. The results showed no significant difference between the effects of the two task types on vocabulary learning and thus it partially supported the involvement load hypothesis as it revealed that the task with low involvement led to more vocabulary knowledge than the task with high involvement.

In a meta-analysis, Yanagisawa and Webb (2021) examined the general predictive ability of the ILH, the relative effects of need, search, and evaluation, and the impact of potential factors such as time on task, frequency of encounters or use, and test format as moderating learning. In doing so, they explored 398 effect sizes from 42 empirical studies. Their results revealed that the ILH was a significant predictor of learning and explained 15.0% of the variance in effect sizes on immediate and posttests and 5.1% of the variance in effect sizes on delayed posttests. Furthermore, they found that the evaluation component was the greatest contributor to learning, followed by need. Conversely, their findings indicated that search did not contribute to learning. Their analysis of moderators demonstrated that test format and frequency moderated learning gains. In another meta-analysis conducted by Huang et al. (2012), it was revealed that learners who performed a task with a higher degree of involvement load, gained more vocabulary.

Some researchers have found counter-evidence to the involvement load hypothesis. For example, Folse’s (2006) study revealed that participants doing three cloze exercises (need+, search-, evaluation+) outperformed those doing one sentence-making exercise (need+, search-, evaluation++). Similar results were obtained by Lu (2013). Lafrer (2003) also questioned this hypothesis as she found that sentence-completion plus dictionary use (need+, search-, evaluation+) led to better vocabulary learning than sentence-writing (need+, search-, evaluation++). Folse (2006), Keating (2008), and Webb (2005) observed that the superiority of the tasks with higher involvement indices disappeared when equal amount of time was assigned to the tasks. Time on task has been found to bear relevance to the predictive power of ILH. For instance, Yanagisawa and Webb (2021) found that involvement load was more contributive to learning compared to time on task. Huang et al.’s (2012) results revealed that time on task had positive effects on vocabulary learning. Folse’s (2006) study showed that when the time was controlled, the task with higher involvement indexes could not lead to better retention. Therefore, as Nation ISP and Webb SA (2011) note time can make the expectations of ILH inverse because the learners might obtain better results if they spend more time on the set tasks. Similarly, as Gohar et al. (2018) contend the time needed for elaborate thinking is an important reason when it comes to task performance as more time spent on a task leads to better vocabulary learning gains.

The present study is carried out in line with previous investigations to shed more light on an underexplored aspect of Involvement Load Hypothesis. As Lafrer and Hulstijn (2001) maintain, based on ILH tasks with varying degrees of involvement loads result in different amounts of learning. However, the number of studies examining if tasks with similar loads may contribute
differently to vocabulary acquisition and retention is scant. Apart from enriching the literature on Involvement Load Hypotheses, the findings of the current study are likely to provide insights for teachers as to what tasks can be more conducive to vocabulary learning and should then be incorporated into their classroom practice.

3. The current study
A review of the previous empirical studies indicates that while some studies on involvement load hypothesis have investigated tasks with different loads of involvement (e.g., Keating, 2008; Lu, 2013; Nassaji & Hu, 2012), others have focused on those tasks with the same loads yet different degrees of prominence (e.g., Folse, 2006), but little attention has been paid to similar tasks regarding their loads, as well as degrees of prominence. As Kim (2008) contends, future studies with a focus on degrees of prominence are required to test the involvement load hypothesis more precisely. Therefore, any study that investigates such tasks is essential to have a better understanding of the ILH and its predictive power. Moreover, to the best knowledge of the researchers, the number of investigations studying the effects of sentence-making, summarizing stories, and making predictions about the end of incomplete stories with similar involvement loads on EFL learners’ acquisition and retention of vocabulary is scant. In fact, the main reason behind the choice of the tasks in the current study is that the literature is replete with studies on composition writing and sentence making (e.g., Folse, 2006; Hulstijn & Lauffer, 2001; Kim, 2008; Lauffer, 2003) but investigations on other modes of written output i.e. summarizing and predicting the end of the story tasks are scant which will be the focus of the present study. Furthermore, as Laufer and Hulstijn (2001) noted, more effective tasks have higher involvement load and are more conducive to learning. Given that the three task types selected for the current study have not been investigated previously in terms of their contributions to vocabulary learning, the findings of the current study can shed light on which one of the three tasks might be more effective. Accordingly, this study aimed to investigate the extent to which three output activities are conductive to learners’ acquisition and retention of the target words. As stated by Hulstijn and Lauffer (2001), the three tasks used in the present study including (1) sentence-making, and two modes of composition writing- (2) summarizing stories, and (3) making predictions about the end of incomplete stories all had an involvement load of 3. According to Hulstijn and Lauffer (2001), the three tasks have a moderate index of one for need (need+) as learners are asked by the teacher to do the activity, a lack of index for search (search-) since learners do not have to use dictionaries or other resources to do the activity, and an evaluation index of two (evaluation++) since learners have to identify the meanings of the target words to do the tasks and also evaluate whether they have used the right words in their contexts of use. In line with the purposes of the present study, the following research questions were formulated:

(1) Do sentence-making, summarizing, and predicting the end of an incomplete story using the target words have any significant effect on the acquisition and retention of those words?

(2) Do different tasks with the same loads of involvement lead to significant differences in the acquisition and retention of target words?

4. Method

4.1. Design
The design adopted in the present study was quasi-experimental as pure random sampling was not feasible (Best & Kahn, 2006). The sampling system of the current study was cluster sampling, since it was not possible for the researchers to randomly assign the participants to three groups. There were three intact classes which were randomly assigned to one of the three experimental conditions.
4.2. Participants and setting
The current study was carried out in a language institute in Tehran, Iran. The classes at the institute contained approximately 28 learners each and learners attended general English classes. Prior to attending the classes all learners had either taken a placement test (four skills included) administered by the institute or passed all the previous level final exams. The participants were 78 EFL learners within the age range of 15 to 18. Most (N = 57) had been studying English for three years and some (N = 21) had passed the placement test of the institute and all were at the intermediate level. They came from three classes and all of them attended the class twice a week for one and a half hours each session. Each group was randomly assigned to one of the experimental conditions. Moreover, the participants’ own teachers assisted the researchers during the treatment sessions and the data collection phase.

To observe the ethical considerations for the current study, initially the participants were provided with a consent form. They were also asked to take the consent form to their parents as the participants were within the age range of 15 to 18. In the consent form, the purposes of the study and data collection had been explained. Moreover, there was information concerning the confidentiality of the test results as well as voluntary participation of students. The researchers requested the parents to sign the consent form and return it to the researchers if they approved the conditions.

4.3. Target vocabulary
In order to select the target words, a sample population of 20 intermediate students who were not supposed to attend the main study were given two reading comprehension texts, the same texts that were used in the main study, and were required to underline the unknown words. The initial 32 words that were selected by the students in the piloting session were included in the pretest. During the pretest phase, 15 out of 32 words which were least known by the participants were selected as target words. These were the words for which most of the participants could not write correct L1 translations or English definitions. The target words included eight nouns, four verbs, two adjectives, and one adverb: appointment, intersection, sight, cane, clue, to overcome, handicap, to interrupt, attitude, naughtiness, sincere, to misbehave, novice, to occur, deliberately.

4.4. Materials
a Reading comprehension texts: This study used two reading passages adapted from Chicken Soup for the Soul (Canfield et al., 2007). The passages were carefully chosen by the researchers so that they were suitable for all the participants in the three experimental groups. The first passage included 269 words and contained 7 target words, and the second passage included 215 words which contained 8 target words. These target words were underlined and bolded in the text to catch the learners’ attention (Schmidt, 1994) with the meanings provided for the learners in parentheses following the words. Moreover, the texts included only one single occurrence of the target vocabulary items to avoid multiple exposure effects (Laure & Hulstijn, 2001). To calculate level of text difficulty, a readability scale provided by a free online tool called Textalyser (http://textalyser.net/) was employed. The Textalyser calculates readability index according to the procedures proposed by Gunning (1952) as follows:

(1) Counting the exact number of words and sentences
(2) Dividing the total number of words by the total number of sentences to arrive at the average sentence length (ASL)
(3) Counting the number of words with three or more syllables
(4) Dividing the number gained in step three by the number of words in the passage (Percentage of hard words: PHW)
(5) Adding the ASL and PHW amounts obtained in steps three and four respectively
(6) Multiplying the result by 0.04
According to Gunning (1652), a readability index of around 11 or 12 is considered very difficult, a readability index between 5 to 7 is regarded moderate and a readability index lower than 5 is interpreted easy. The readability indices for the texts in the present study were all within the range of 5 to 7 (M = 5.5) which is considered appropriate for the intermediate level learners.

b Vocabulary learning activities: In this study, three output activities,—namely, sentence-making, summarizing a text using the target words, and predicting the end of a story- were used. According to the involvement load hypothesis, the involvement indices of all the above activities are 3 (need+, search-, evaluation++). Need is moderate in all of the activities because the they imposed learners to know the meanings of the words. The search component is absent in all the three activities since the meanings of the target words were provided for the learners in the reading texts. As for evaluation, it is strong because the participants had to decide how to fit the words in a specific context in all three conditions. Based on the predictions of the involvement load hypothesis, these tasks are all equally effective in vocabulary learning, since they all induce the same loads of involvement.

4.5. Instruments
In the current study, following researchers such as Hulstijn and Laufer (2001), Keating (2008), and Zou (2017), a translation test was used to measure the learners’ acquisition and retention of the target words. According to Nation ISP (2001), a translation test may be a more effective method to measure the knowledge of the meaning than a multiple-choice test. Participants were provided with a list of target words and were required to provide the L1 (Persian) translations or the L2 (English) synonyms or definitions of the target words. The order of the target words was changed on the immediate and delayed posttests. Following Hulstijn and Laufer (2001), Keating (2008), and Zou (2017), the scoring system of the translation test comprised either 0 or 1 for each target word. If the participants could not supply the meaning of the target word, 0 was awarded to that word. The response was graded as 1 when the learners could give a correct English synonym or L1 translation. Therefore, the maximum score was 15 points.

4.6. Procedure
The procedure of this study was composed of two main phases: piloting and the main phase. During the piloting session, a sample of 20 students who were at the same level with the participants in the main study were required to read the two passages and underline the unfamiliar words. These words were later included in the pretest.

The main phase of the study included three sessions- pretest, reading the texts, and doing the output activities (treatment) plus one immediate posttest, and one delayed posttest. One week before the treatment session, the pretest was administered that included 15 words. During the second session, which included treatment and the immediate posttest, the learners were instructed to perform a specific activity based on their grouping.

First, participants in all the groups were given the two reading texts and 15 minutes to read them, and were told in advance about the task they were required to do following the reading. Then, they were asked to do the activities based on their grouping. Moreover, following Laufer and Hulstijn (2001), participants were not told that they were going to be tested on the recall of the target words in advance. Furthermore, like Hulstijn and Laufer (2001) study, time-on-task was considered as the inherent property of the tasks in the current study and accordingly the learners did not have a time-limit to complete the tasks in any of the groups. However, time-on-task was recorded for the groups to help analyze the prospective results in the light of time-on-task.

The procedure for each group is detailed below:
4.6.1. Sentence--making

After reading, the texts were collected and the participants were given a worksheet on which the target vocabulary items with their L1 translations along with their parts of speech were available. The participants were asked to write 15 sentences using the target words and also to pay attention to both grammar and meaning appropriateness of their sentences. Moreover, they were told not to use the same sentences of the reading texts. The time was recorded by the one of the researchers. The participants returned their papers in almost 30 minutes.

4.6.2. Summarizing the texts

Students in this group were asked to read the two texts carefully because they had to summarize the stories in their absence. After collecting the texts, the participants were each given two worksheets on which there were the target vocabulary items with their L1 translations along with their parts of speech. The participants were required to summarize both reading passages, each in one paragraph (7-8 lines), using the target words. They were also told to pay attention to the coherency of the paragraphs and try to use the words correctly in the context. It took almost 45 minutes for the participants to finish the task.

4.6.3. Predicting the end of the stories

Participants in this group were told that they were required to read two incomplete stories and use what they already knew to predict the end of the stories and write them down, each in one paragraph. Following reading the texts, the learners were provided with worksheets on which the target words with their L1 translations and parts of speech were available. The participants returned their papers in almost 50 minutes.

Following the treatment, all participants received the immediate posttest to assess the extent to which the activities contributed to the acquisition of the words. Two weeks later, another posttest was administered.

5. Results

5.1. Pretest results

Initially, it was needed to make sure that the three groups of the study were not significantly different in terms of vocabulary knowledge. This was necessary, otherwise the differences among the groups on the posttest could not be attributed to treatment types. To statistically probe any significant differences among the three groups on vocabulary pretest, a One-way ANOVA was run on the pretest scores of the three groups. Table 1 presents the results of descriptive statistics for the vocabulary pretest scores.

Before running ANOVA, it was needed to check the assumption of homogeneity of variances. Table 2 displays the results for the Test of Homogeneity of Variances run on the pretest scores for the three groups.

As indicated in Table 2, all the sig values are above .05. Therefore, it can be inferred that the assumption for Homogeneity of Variances is met and thus running ANOVA is guaranteed. Table 3 illustrates the results of ANOVA run on the pretest scores for the three groups.

As shown in Table 3, the sig value is above 0.05 and therefore it can be concluded that there were no significant differences among the three groups on the pretest of vocabulary.

5.2. Addressing the first research question

The first research question asked whether sentence-making, summarizing, and predicting the end of a story using the target words had any significant effects on the acquisition and retention of those words. To answer this question, descriptive statistics (Table 4) for the participants’ scores in pretest, immediate, and delayed posttest were computed.
As Table 4 demonstrates, there was an increase in the means from pretest to immediate posttest for all the groups.

To measure the significance of the mean rise from pretest to immediate posttest, Wilcoxon Signed Ranks Test as a non-parametric test was run. Table 5 presents the Wilcoxon Signed Ranks Test results.

As presented in Table 5, the sig value for the sentence-making group equals .00 which is lower than the critical value (p < .01). Similarly, the sig value for the summarizing and prediction groups both equal .00 which are lower than the critical value (p < .01). Thus, it can be inferred that the three task types had significant effects on the acquisition of target words.

In order to measure the significance of the mean decline from immediate posttest to delayed posttest in the three groups, three paired samples t-Test as a parametric test were run. Table 6 illustrates the paired samples t-Test results.

As depicted in Table 6, the sig values for the three groups are lower than the critical value (p <.01). Therefore, it can be inferred that the three task types did not have any significant effects on the retention of target words. Thus, sentence-making, summarizing, and predicting the end of an incomplete story using the target words had significant effects on the acquisition of the target words but not on the retention of those words.

5.3. Addressing the second research question
The second research question asked whether the tasks with the same loads of involvement lead to significant differences in the acquisition and retention of the target words. In order to investigate the effects of the three word-focused tasks on vocabulary acquisition, an ANCOVA test along with post hoc analysis was used because the descriptive statistics showed that the groups were not equal to each other on the pretest: participants in the predicting group had the highest mean (.2963), followed by the summarizing group (.2800). Table 7, showing the adjusted immediate posttest means after controlling the covariate effect by ANCOVA, indicates that the predicting group had the highest immediate posttest mean (9.844), and the sentence-making group had the lowest mean (8.271).

The results of the ANCOVA test are reported in Table 8. As can be seen, the groups were significantly different on the immediate posttest; F (2,74) = 7.167, p < .05; however, it was not clear which group had the highest gain in the acquisition of target words.

Since the study involved three groups, post hoc pairwise comparisons were run to find out which of these groups had significantly higher and lower means (Table 9). The results indicated that the predicting group with the highest mean significantly outperformed the sentence-making and summarizing groups (p < .01), while no statistical difference was found between the
sentence-making and the summarizing groups. Concerning the involvement load hypothesis, the results were partially consistent with its predictions, as the effectiveness of sentence-making and summarizing showed no statistically significant difference. However, the predicting group was found to be more effective than the other two tasks, although they were all accorded equal loads.

Regarding the delayed posttest, there were significant declines for all the groups from the immediate to the delayed posttest. That is, none of the tasks significantly affected vocabulary retention. Therefore, attempt was made to identify which method resulted in most or least decline in the knowledge of target words. As can be seen in Table 4, the groups were unequal to each other in the immediate posttest: sentence-making group with the mean of 8.2308, summarizing group with the mean of 8.4000, and predicting group with the mean of 9.8889. Therefore, ANCOVA test was used to control the effect of initial differences (i.e., immediate posttest). As Table 10 demonstrates, there was a significant difference among the groups in delayed posttest; F (2,74) = 11.926, p < .05.

These differences are better illustrated in the adjusted delayed posttest means (Table 11) after controlling the covariate effect. The results of ANCOVA showed that the summarizing group had the highest mean (6.325), followed by the predicting group (6.207), while the sentence-making group obtained the lowest mean (5.318).

In order to locate the differences, Bonferroni’s post hoc pairwise comparison was run (Table 12) which indicated that both predicting and summarizing groups performed significantly better than the sentence-making group (p < .01). Furthermore, the predicting and summarizing groups were not significantly different. It is concluded that although participants in all the three groups did not maintain the target words knowledge on the delayed posttests, the sentence-making group showed the most decline on the delayed posttest, that is to say it was the least helpful in maintaining target words knowledge. The summarizing and predicting groups, however, showed the same amount of decline on the delayed posttest.

Therefore, it can be concluded that the participants in the predicting group outperformed those who completed the other two tasks on delayed posttest. That is, the predicting group had the

| Table 2. The Results of Homogeneity of Variances Test for the Vocabulary Pretest Scores |
|---------------------------------|-----------------|-----|-----|-----|
| Pretest Scores                  | Levene Statistic | df1 | df2 | Sig. |
| Based on Mean                   | .010            | 2   | 75  | .990 |
| Based on Median                 | .016            | 2   | 75  | .985 |
| Based on Median and with adjusted df | .016       | 2   | 74.169 | .985 |
| Based on trimmed mean           | .025            | 2   | 75  | .975 |

| Table 3. The Results of ANOVA for the Vocabulary Pretest Scores of the Three Groups |
|---------------------------------|-----------------|-----|-----|-----|
| Pretest Scores                  | Sum of Squares  | df  | Mean Square | F   | Sig. |
| Between Groups                  | .010            | 2   | .005        | .016| .985 |
| Within Groups                   | 23.785          | 75  | .317        |     |     |
| Total                           | 23.795          | 77  |             |     |     |
highest mean, followed by summarizing group, with the lowest mean belonging to the sentence-making group.

6. Discussion
The aim of this study was to examine the effects of three tasks, namely- sentence-making, summarizing a text, and predicting the end of an incomplete story-on vocabulary acquisition

| Group       | Mean      | Std. Error | 95% Confidence Interval |
|-------------|-----------|------------|-------------------------|
| sentence-making | 8.271a    | .329       | 7.616 to 8.926          |
| summarizing  | 8.406a    | .335       | 7.738 to 9.075          |
| predicting   | 9.844a    | .323       | 9.201 to 10.487         |

a. Covariates appearing in the model are evaluated at the following values: Pretest = .282.
and retention in the light of involvement load hypothesis. The first question asked if the tasks of making sentences, summarizing, and predicting the end of an incomplete story using the target words would promote learners' acquisition and retention of those words. The results indicated that these tasks were effective in improving learners' acquisition of the words but not on the retention of those words. The findings in terms of the effects of output tasks on vocabulary acquisition are consistent with a number of previous studies (e.g., Keating, 2008; Laufer, 2003; Webb, 2005). These studies have indicated that written output tasks are contributive to vocabulary acquisition. Regarding vocabulary retention, the results of the current study are similar to the findings from some other studies such as Keating (2008), Pichette et al. (2012), and Lu (2013), who found that the superiority of output tasks disappear over time.

The first plausible explanation for the significant decline between immediate and delayed posttest is the design of the current study in which there was no additional exposure to the target words and no repeated opportunities for the learners to retrieve them. A number of studies have pointed to the importance of repeated retrieval opportunities when vocabulary learning is considered (e.g., Elley, 1989, as cited in Nation ISP, 2001). According to Gathercole
and Baddeley (1990, as cited in Nation ISP, 2001), each time a word is retrieved, the link between its form and meaning gets stronger in the mind. Research shows that learners can retrieve the meanings of recently learned words several weeks after the first meetings; however, the length of the time depends on the number of previous meetings (Nation ISP, 2001). This explanation is also in accordance with what Hulstijn and Laufer (2001, pp. 274, footnote 20) argued: “one expects a decline in knowledge over time in the absence of rehearsal or additional exposure to the target words”. Besides, as Nation ISP (2001) stated, learners seem to forget most of the recently learned items after they are first studied. Thus, he further suggested that new items should be repeated very soon after initial learning. As Anderson and Jordan (1928, as cited in Nation) observed, only 39% of the material were retained by the participants in their study after three weeks. Therefore, the results obtained in the current study do not seem surprising.

Another explanation might be that the learners were provided with the meanings of the words and were not required to look them up, that is, the search component was absent in all the tasks. The importance of search as a factor in vocabulary learning has been highlighted by several researchers. For instance, Hulstijn (1992), Cho and Krashen (1994), and Laufer (2003) pinpointed the superiority of the tasks which required the learners to look up the words in a dictionary.

The last justification for the significant decline between immediate and delayed posttests might be related to the testing instrument used in the current study. In order to measure participants' receptive knowledge of the words, translation tests were employed in both immediate and delayed posttest. However, in the field of vocabulary, most studies have used multiple-choice items to measure learners’ receptive knowledge of the words. According to Nation ISP (2001), these tests are much easier for learners compared with translation tests, and learners need “strong knowledge of

| Table 10. Tests of Between-Subjects Effects on Delayed Posttest Means |
| Dependent Variable: Delayed.Post |
| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
| Corrected Model | 586.219² | 3 | 195.406 | 303.936 | .000 | .925 |
| Intercept | 61.455 | 1 | 61.455 | 95.588 | .000 | .564 |
| Group * Immediate. Post | .508 | 2 | .254 | .389 | .679 |
| Immediate.Post | 493.298 | 1 | 493.298 | 767.279 | .000 | .912 |
| Group | 15.335 | 2 | 7.668 | 11.926 | .000 | .244 |
| Error | 47.576 | 74 | .643 |
| Total | 3394.000 | 78 |
| Corrected Total | 633.795 | 77 |

| Table 11. Adjusted Delayed Posttest Means after Controlling the Covariate Effect by ANCOVA |
| Dependent Variable: Delayed.Post |
| Group | Mean | Std. Error | 95% Confidence Interval |
| | | | Lower Bound | Upper Bound |
| sentence-making | 5.318² | .159 | 5.001 | 5.635 |
| summarizing | 6.325² | .161 | 6.004 | 6.647 |
| predicting | 6.207² | .159 | 5.890 | 6.525 |

| a. Covariates appearing in the model are evaluated at the following values: Immediate.Post = 8.8590. |
Table 12. Pairwise Comparisons of Delayed Posttest Means

| Dependent Variable: Delayed_Post | (I) Group | (J) Group Mean Difference (I-J) | Std. Error Sig. | b) Adjust for multiple comparisons: Bonferroni |
|-----------------------------------|-----------|---------------------------------|-----------------|-----------------------------------------------|
| sentence-making summarizing       | .107*     | .225 .000                        |                 |                                               |
| predicting                        | -.889*    | .229 .001                        |                 |                                               |
| summarizing sentence-making       | .118      | .230 1.000                       |                 |                                               |
| predicting                        | .107*     | .225 .000                        |                 |                                               |
| summarizing sentence-making       | -.889*    | .229 .001                        |                 |                                               |

Based on estimated marginal means
* The mean difference is significant at the .05 level.
b. Adjustment for multiple comparisons: Bonferroni.

The words” for translation tests, while “partial knowledge” is needed for multiple-choice items. (p. 586). He further notes that multiple-choice items might provide opportunities for the learners to guess the correct answers. It can be concluded that the nature of the testing instrument used in this study affected the number of the words the participants could answer correctly, especially in the delayed posttest which was conducted two weeks after the treatment session. If a multiple-choice test were used instead of a translation test, different results might have been obtained.

The second research question asked if tasks with the same loads of involvement result in significant differences in the acquisition and retention of the target words. The results of the immediate posttest revealed that participants in the predicting group outperformed those who completed the other two tasks. That is, the predicting group had the highest mean, followed by summarizing group, with the lowest mean belonging to the sentence-making group. The insignificant difference between the summarizing and sentence-making group provides support for the ILH in that tasks with the same loads of involvement are equally effective in vocabulary acquisition. However, significant differences were observed between the predicting group and both summarizing and sentence-making groups. Consequently, the results were partially consistent with the predictions of the ILH.

The superiority of the predicting group over the other two experimental groups can be explained with respect to the creative nature of this task. The effectiveness of the writing tasks with creative nature in language learning have been pointed out by various researchers. For example, Kenny (2011) claimed that due to using their imaginations in creative writing, learners become more motivated which results in faster and more successful learning. Moreover, what differentiates creative writing from other forms of writing is that in creative writing learners form a link between their pre-existing knowledge to a new linguistic and situational context. This is a process that cannot be underestimated in language learning (Kenny, 2011). This finding is also in line with Rassaei (2017) who observed that participants in the predicting group yielded better results compared with those in other writing groups. Therefore, it seems that owing to its creative nature, the predicting task induces stronger evaluation compared to the other two output tasks.

The results of the study concerning the outperformance of predicting group compared with the other two experimental groups is incongruent with the tenets of ILH since it allocates equal evaluation to all the activities in which the learners are required to generate the context for the new words. Moreover, the predicting task was the most time-consuming compared to the other two tasks. The average completion time of the predicting task was almost 50 minutes, while it took almost 45 and 30 minutes respectively for participants in the summarizing and sentence-making groups to complete the tasks. Thus, it seems plausible to argue that time-on-task is an important factor compared to degrees of prominence as the results of the current study indicated that although the three tasks had similar ILH indices, learners’ performances in the three groups
were different. Since the predicting group obtained better results compared to the other two groups and as the participants in this group outperformed the other two, it is safe to argue that time-one-task can be a more plausible predictor for vocabulary learning compared to degrees of prominence. As Nation ISP and Webb SA (2011) argue, need, search, and evaluation are not the only factors in determining task effectiveness, and there may be a need to add more factors to ILH.

With regard to time-one-task, Hulstijn and Laufer (2001) note that, more demanding tasks need more time to complete. Consequently, it can be speculated that the predicting task in this study was the most demanding for the participants. Students in the predicting group, like the sentence-making group, had to evaluate the target words against the other words in a sentence, check the syntax and collocations of the sentence, pay attention to semantic correction, and like the summarizing group decide how to connect the sentences and make a coherent text. Unlike the participants in the other two groups, though, the participants in the predicting group had to create a sequence of actions imagined. However, concerning the summarizing and sentence-making groups, the results lend support to the involvement load hypothesis. This finding is in agreement with the results obtained in Kim’s (2008) second experiment in which sentence-making and writing a composition using the target words enhanced learners’ vocabulary knowledge equally. All in all, it seems that there was only the creative nature of the predicting task that resulted in a significant difference between this task and the other two.

Regarding the delayed posttest, none of the tasks in the current study could contribute significantly to the retention of the target words. Investigating the results to indicate which task led to the least decline from the immediate to the delayed posttests revealed that unlike the predictions of the ILH, participants completing the two composition writing tasks, i.e. summarizing and predicting, performed remarkably better than the sentence-making group. One plausible explanation concerning the superiority of the two composition-writing tasks over sentence-making is the degree of pre-task planning. Participants in all the three groups were required to generate the context for the target words. Prior to that, they needed to practice how to use the target words in the context in their minds during pre-task planning. Composition-writing tasks of any kind induce a more demanding pre-task planning, since the learners have to generate individual sentences in which the target words could fit and connect the sentences in a coherent text (Zou, 2017).

7. Conclusions and implications
The results of the current study revealed that the three tasks significantly affected vocabulary learning. However, it was found that the three tasks with the same loads of involvement led to significant differences in the learning and retention of the target words. More specifically, the results indicated that the predicting group with the highest mean significantly outperformed the sentence-making and summarizing groups with respect to vocabulary learning. As for vocabulary retention, it was demonstrated that both predicting and summarizing groups performed significantly better than the sentence-making group. Overall, it was concluded that the findings of the current study partially support the ILH, that is, tasks with identical involvement loads are not always equally contributive to the acquisition and retention of the words. Therefore, a significant conclusion drawn from the results of this study is that, some factors such as the number of the times a task exposes the learners to the target words and requires them to retrieve the words, and the presence or absence of the components of the ILH, i.e. need, search, and evaluation as well as the time-on-task factor play a determining role in task effectiveness. Therefore, further research would be beneficial in order to elaborate and refine the evaluation component of ILH.

Based on the findings of the present study, EFL teachers should employ prediction tasks more in comparison with sentence-making and summarizing tasks since the findings revealed that the predicting group with the highest mean significantly outperformed the sentence-making and summarizing groups regarding vocabulary learning. With regard to vocabulary retention, teachers should employ predicting and summarizing tasks more instead of sentence-making tasks since the
findings indicated that both predicting and summarizing groups performed significantly better than the sentence-making group with regard to vocabulary retention.

This study like previous studies in the field of SLA was prone to some limitations. First, participants were all at the intermediate level, and their ages ranged from 15 to 18. Replicating this study, but with participants at different levels of proficiency and age range could result in obtaining different findings. Second, following Hulstijn and Lafer (2001), time-on-task was considered as an inherent property of a task. As stated before, some researchers believe that when tasks are compared, the same amount of time should be allocated to all of them. So, studies can be conducted in which the same time length is assigned to all the tasks in order to reveal the effect of time on vocabulary acquisition. Third, this study could benefit from introspective data collection procedures, such as think-aloud and interview, to gain better insights on the process of vocabulary learning. Replicating this study employing one of these procedures would enable researchers to have better interpretations on the results of the study. Furthermore, in order to measure learners’ receptive knowledge of the words, a translation test was used in the current study. Other testing instruments such as multiple-choice tests or tests that measure learners’ knowledge of the words in a more communicative environment could also be used. Finally, measuring productive knowledge of the words was not included in this study due to time limitations. Thus, measuring both aspects of word knowledge, i.e. receptive and productive, seems necessary and warrants further research.

Funding
The authors received no direct funding for this research.

Author details
Sahar Taheri1
E-mail: taheri.sahar89@gmail.com
Ghafour Rezaie Golandouz2
1 Department of English Language and Literature, Science and Research Branch, Islamic Azad University, Tehran, Iran.
2 Department of English, Islamic Azad University, Garmso, Iran.

Citation information
Cite this article as: The effect of task type on EFL learners’ acquisition and retention of vocabulary: an evaluation of the involvement load hypothesis, Sahar Taheri & Ghafour Rezaie Golandouz, Cogent Education (2021), 8: 1915226.

References
Amini, A., & Maftoon, P. (2017). The impact of skill integration on task involvement load. The Journal of English Language Pedagogy and Practice, 10(21), 29-48.
Best, J. W., & Kohn, J. V. (2006). Research in education. Boston. Pearson Education, Inc.
Confield, J., Mv, H., & Unkovitch, A. (2007). Chicken soup for the soul in the classroom. Health Communications.
Cho, K. S., & Krashen, S. (1994). Acquisition of vocabulary from the sweet valley kids series: adult EFL acquisition. Journal of Reading, 37, 662–667. http://www.sdkrashen.com/content/articles/1994_sweet_garden_with_sweet.pdf
Craik, F. 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/10.1016/0022-5371(72)80001-X
Elley, W. B. (1989). Vocabulary acquisition from listening to stories. Reading Research Quarterly, 24(2),174-187. https://doi.org/10.2307/74863
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
Gathercole, S. E., & Boddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. British Journal of Psychology, 81(1), 439–454. https://doi.org/10.1111/j.2044-8295.1990.tb02371.x
Gohar, M. J., Rahamanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. Journal of Psycholinguistic Research, 47(4), 859–869. https://doi.org/10.1007/s10936-018-9568-5
Gunning, R. (1952). The technique of clear writing. McGraw-Hill.
Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clock organ: Acquiring second language vocabulary through reading. Reading in a Foreign Language, 11(2), 207–223. https://bupevu.ga/part_223.pdf
Huang, S., Willson, V., & Estami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. The Modern Language Journal, 96(4), 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x
Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental learning. In P. J. Arnaud & H. Bejoint (Eds.), Vocabulary in Applied Linguistics (pp. 113–125). Macmillan.
Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. The Modern Language Journal, 80(3), 327–339. https://doi.org/10.1111/j.1540-4781.1996.tb01614.x
Hulstijn, J. H., & Lafer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. Language Learning, 51(3), 539–558. https://doi.org/10.1111/1042-4431.00164
Keating, G. (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/10.1177/136216880809922
Kenny, S. (2011). Teaching creative writing in an ESL context. OTB Forum, 4(1), 50–54. http://otbforum.net/otb_forum/4-1/kenny_2011_otb_forum_4-1.pdf

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. Language Learning, 58(2), 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Krashen, S. (1989). We acquire vocabulary and spelling by reading: additional evidence for the input hypothesis. The Modern Language Journal, 73(4), 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.x

Lafer, 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

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

Lafer, B., & Rozovski-Raibbot, 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/10.1177/13621688114212019

Lu, M. (2013). Effects of four vocabulary exercises on facilitating learning vocabulary y meaning, form, and use. TESOL Quarterly, 47(1), 167–176. https://doi.org/10.1002/tesq.79

Nomazidianost, E., Hosseini, E., & Wahyu, U. D. (2020). A comparative effect of high involvement load versus lack of involvement load on vocabulary learning among Iranian sophomore EFL learners. Cogent Arts & Humanities, 7(1), 1. https://doi.org/10.1080/23311983.2020.1715525

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 ISP. (2001). Learning vocabulary in another language. Cambridge University Press.

Nation ISP and Webb SA. (2011). Researching and analyzing vocabulary. Heinle.

Newton, J. (2001). Options for vocabulary learning through communication tasks. English Language Teaching Journal, 55(1), 30–37. https://doi.org/10.1080/00222130110013055.1.30

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

Peters, E., Hulstijn, J. H., Serca, I., & Lutjensmars, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. Language Learning, 59(1), 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Pichette, F., De Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. Applied Linguistics, 33(1), 66–82. https://doi.org/10.1093/applin/amr037

Pourkabiri, A. A., & Biria, R. (2015). Efficacy of task-induced involvement in incidental lexical development of Iranian senior EFL students. English Language Teaching, 8(5), 122–131. https://doi.org/10.5539/elt.v8n5p122

Rahmani, R., Jafari, S., & Izadpanah, S. (2018). The effect of task induced involvement load on unfamiliar L2 vocabulary learning: sentence writing, summary writing, imaginary story writing and creative sentence writing. Applied Research on English Language, 7(1), 67–88. https://www.ssid.irdien/journal/ViewPaper.aspx?id=693197

Rassaei, E. (2017). Effects of three forms of reading-based output activity on L2 vocabulary learning. Language Teaching Research, 21(1), 76–95. https://doi.org/10.1077/1362168816630160

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

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. Ellis (Ed.), Implicit and Explicit Learning of Languages (pp. 165–209). Academic Press.

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

Swain, M. (1993). The output hypothesis: Just speaking and writing aren’t enough. Canadian Modern Language Review, 50(1), 158–164. https://doi.org/10.3138/cmlr.50.1.158

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 471–483). Lawrence Erlbaum Associates.

Un-udom, N. (2018). The effects of task-induced involvement on EFL vocabulary learning. Chophayom Journal, 29(3), 181–196. https://scopus.sci-thaijo.org/index.php/ejChophayom/article/view/154925

Webb, S. (2005). Receptive and productive vocabulary learning: the effects of reading and writing on word knowledge. Studies in Second Language Acquisition, 27(01), 33–52. https://doi.org/10.1017/S02772263050050023

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

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. Language Learning, 71(1), 121–146. https://doi.org/10.1111/lang.12444

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), 1–22. https://doi.org/10.1177/1362168816652418
Appendix 1
The first passage used in the current study

He was in the third grade class I taught at Saint Mary’s School in Morris, Minnesota. All 34 of my students were dear to me, but Mark Eklund was one in a million. Very neat in appearance, he had that positive attitude that made his occasional naughtiness lovely.

Mark also talked all the time. I tried to remind him again and again that talking without permission was not acceptable. What impressed me so much, though, was the sincere response every time I had to correct him for misbehaving (to misbehave: نابضیاری کردن). “Thank you for correcting me, Sister!” I didn’t know what to make of it at first but before long I became accustomed to hearing it many times a day.

One morning I was losing patience with him when Mark talked once too often. I made a novice (تفریکری کردن)–teacher’s mistake. I looked at Mark and said, “If you say one more word, I am going to tape your mouth shut!”

It wasn’t ten seconds later when one of the students suddenly said “Mark is talking again.” I hadn’t asked any of the students to help me watch Mark, but because I had stated the punishment in front of the class, I had to act on it.

I remember the scene as if it had occurred (to occur: افتادن افتادگی) this morning. I walked to my desk, very deliberately (بی‌پایان بی‌پایان) opened the drawer and took out a roll of masking tape. Without saying a word, I moved toward Mark’s desk, tore off two pieces of tape and made a big X with them over his mouth. I then returned to the front of the room . . .

Appendix 2
Sample Tasks

Task 1: Sentence-making

Applying از ابتدای اعمال کودکان از ابتدای اعمال کودکان از ابتدای اعمال کودکان از ابتدای اعمال کودکان از ابتدای اعمال کودکان.

Attitude (n): رفتار
Naughtiness (n): شکطات
Sincere (adj.): صادقانه
To misbehave (v): نابضیاری کردن
Novice (adj.): نابهاری کار
To occur (v): افتادن افتادگی
Deliberately (adv.): غیر
Appointment (n): قراردادهای
Intersection (n): چهارراه
Sight (n): منظره
Cane (n): عصا
To overcome (v): غلبه کردن
Handicap (n): معلولیت
Clue (n): سریع

Task 2: Summarizing (Reading No. 1)

He was in the third grade class ...

Task 3: Predicting (Reading No.1)

He was in the third grade class ...

………………
