Form-focused Instruction in Second Language Vocabulary Learning: A Case for Contrastive Analysis and Translation (A Conceptual Replication Study)[1]

Ali Jahangard[2], Languages and Linguistics Center, Sharif University of Technology, Tehran, Iran

Abstract

One of the most interesting studies on the role of L1 and contrastive analysis in vocabulary teaching is by Laufer and Girsai (2008). However, due to some methodological issues, their research findings are open to criticism and controversy. The current study aimed to replicate the research with a more rigorous design to re-investigate the efficiency of three different methods of teaching new vocabularies to EFL learners. To meet this end, 105 intermediate adult participants were randomly assigned to three groups, each receiving instruction in a specific method. Sixteen target words, later embedded into a reading text, were used as instructional materials. Group 1 received non-contrastive Meaning Focused Instruction (MFI); Group 2 was exposed to non-contrastive Form Focused Instruction (FFI), and Group 3 was instructed using the Contrastive Analysis and Translation (CAT) method. Then, immediate and delayed post-tests were administered to examine the retention of the target words. The tests were designed so that the results were less affected by the compatibility between the instruction and the testing method. Nevertheless, the results indicated that each method lead to differential amounts of retention and that the compatibility between the method of treatment and the testing method might affect the results.

Resumen

Uno de los estudios más interesantes sobre el papel de la L1 y el análisis contrastivo en la enseñanza del vocabulario es el de Laufer y Girsai (2008). Sin embargo, debido a algunos problemas metodológicos, los resultados de dicha investigación están abiertos a críticas y controversias. El estudio actual tuvo como objetivo replicar la investigación con un diseño más riguroso para indagar la eficiencia de tres métodos diferentes de enseñar nuevos vocabularios a los estudiantes de inglés como lengua extranjera. Para este fin, 105 participantes adultos intermedios fueron asignados aleatoriamente a tres grupos intactos, cada uno de los cuales recibió instrucción en un método específico. Dieciséis palabras objetivo que luego se incrustaron en un texto de lectura se usaron como materiales de instrucción. El Grupo 1 recibió Instrucción Centrada en el Significado no contrastiva; el grupo 2 fue expuesto a Instrucción Enfocada en la Forma no contrastiva, y el grupo 3 fue instruido usando el método de Análisis y Traducción Contrastivo. Luego, se administraron pruebas posteriores inmediatas y diferidas para examinar la retención de las palabras objetivo. Las pruebas fueron diseñadas para que los resultados se vieran menos afectados por la compatibilidad entre la instrucción y el método de prueba. No obstante, los resultados indican que cada método conduce a cantidades diferenciales de retención y que la compatibilidad entre el método de tratamiento y el método de prueba podría afectar los resultados.

Introduction

Words are building blocks of languages and without adequate vocabulary little information can be verbally communicated. Second language learners and teachers are consciously or unconsciously aware of the important role that vocabulary plays in effective decoding and encoding of meaning in a second language. However, the question of which learning activities, tasks, or techniques to employ to best enhance vocabulary acquisition has long been an enigma to resolve for many language teachers and researchers. Hence, investigation on the tasks culminating to the most durable vocabulary learning seems to be an important pedagogical necessity. Of course, the intriguing issue of finding the most effective methods of teaching and learning vocabulary has led to an extensive literature, e.g., Schmitt (2008); Wright and Cervetti (2017). However, along with the studies on the comparison of learning activities that produce the most durable word retentions, related research has shown that L1 considerably influences the learning and use of L2 vocabulary in various ways (for a recent literature review, see Puig-Mayenco et al., 2018). As an example, Hemchua and Schmitt (2006) studied the lexical errors in compositions of EFL learners at a Thai university and concluded that about 25% of the errors had roots in their L1. Also, learners consider L1 use or translation a beneficial tool for learning English language skills, such as reading, writing, and especially vocabulary (Liao, 2006). Nevertheless, "Translation in language teaching has been treated as a pariah in almost all the fashionable high-profile language teaching theories of the 20th century" (Cook, 2010. p. xv). Cook also states that, "In the fast growing study of vocabulary acquisition, ... research into translation as a

[1] This is a refereed article. Received: 1 July, 2021. Accepted: 13 December, 2021. Published: 1 September, 2022.
[2] jahangard@sharif.edu, 0000-0002-5319-0456

means of learning is also almost non-existent" (pp. 90-91).

Though the research on the contrastive analysis of learners' L1 and L2 and its effect on the acquisition of L2 morphosyntax is very limited, one of the few pioneering studies regarding this connection was the study by Laufer and Girsai (2008). The initial study examined the effectiveness of three vocabulary instructional methods, i.e., Meaning-Focused Instruction (MFI), Form-Focused Instruction (FFI), and Contrastive Analysis and Translation (CAT). Then, they measured the immediate and delayed retention of the target words and concluded that the contrastive and translation group was significantly superior to the other two groups on all tests. However, though the important role of L1 in the process of L2 acquisition was undeniable, certain methodological issues cast shadows of uncertainty on the initial study's results. Actually, the testing method employed to elicit the related data on vocabulary retention rates is assumed to have affected the study results. In testing the retention rates of the target vocabulary items, they employed translation from L1 into L2 and vice-versa, a testing method compatible with the learning tasks in the highest retention group of the study. This test-task compatibility (using the same format in teaching and testing) arguably, unduly boosted the L1 effects on the retention results in that particular group.

This phenomenon was also referred to as 'transfer appropriateness' by Bransford et al. (1979). There might be compatibility, incompatibility, or neutrality between the retention testing method and the processing mode of the previous learning task (Hulstijn, 2003, pp.356-357). Morris et al. (1977) indicated an interaction between encoding processes and retention tasks (semantic/non-semantics learning vs. semantic/non-semantic retention). In other words, the subjects who were given compatible learning and retention tasks, i.e., semantic vs. semantic, non-semantic vs. non-semantic, gained greater retention scores than those assigned incompatible learning and retention tasks. The instructive point of this research bearing to our argument is that a precise measurement of intentional or incidental learning outcomes requires bilateral regard for learning and retention tasks (Eysenck, 1982). The initial study did not take this point into account. Therefore, it is likely that the better performance of the group which took the compatible retention task was an artifact of the testing method rather than the instructional method. To explain clearly the problem with the study, the results of the initial research are summarized in the table below:

	Learning Tasks	**Retention Tasks**	**Results**	**Learning & Retention Compatibility**
Group 1 Meaning Focused Instruction (MFI)	a. Reading the passage and answering the reading comprehension questions b. Checking the true/false questions c. Providing answers to the reading comprehension questions in the open-ended and fill-in-the-blank format with the text at their disposal d. Paired and grouped discussion of the topics introduced by the teacher. **(the language of the interaction was English only).*	a. (Active Recall) providing the target words in response to their Hebrew translations, i.e., (L1 → L2 translation) b. (Passive Recall) translating the target words into Hebrew, i.e., (L2 → L1 translation)	I. Immediate Post-test: a. Active Recall Single words =0.12 (Max.= 10) Collocations=0.27 (Max.=10) b. Passive Recall: Single words=0.31 (Max.= 10) Collocations= 3.69 (Max.=10) II. Delayed Post-test: a. Active Recall: Single words =0.12 (Max.= 10) Collocations=0.35 (Max.=10) b. Passive Recall: Single words=0.15 (Max.= 10) Collocations=4.54 (Max.=10)	Non-compatible
Group 2 Non-contrastive Form Focused Instruction (FFI)	a. Reading the passage and answering the reading comprehension questions b. checking the true/false questions c. Meaning recognition of the target vocabulary through a multiple-choice exercise d. a text fill-in activity with the target words provided in a word-bank *(the language of the interaction was English only).	a. (Active Recall) providing the target words in response to their Hebrew translations, i.e., (L1 → L2 translation) b. (Passive Recall) translating the target words into Hebrew, i.e., (L2 → L1 translation)	I. Immediate Post-test: a. Active Recall: Single words=2.30 (Max.= 10) Collocations=2.91 (Max.=10) b. Passive Recall: Single words = 4.04 (Max.=10 Collocations = 5.70 (Max.=10) II. Delayed Post-test: a. Active Recall: Single words = 2.13 (Max.= 10) Collocations =3.00 (Max.=10) b. Passive Recall: Single words=3.52 (Max.=10) Collocations= 6.00 (Max.=10)	Non-compatible

Group 3 Contrastive Analysis and Translation (CAT)	a. Reading the passage and answering the reading comprehension questions b. Translating the sentences which included the new target words in the text into Hebrew i.e., (L2 → L1 translation) c. Translating the same sentences from Hebrew into English i.e., (L1 → L2 translation) d. Explicit contrastive analysis and instruction of the new words.	a. (Active Recall) providing the target words in response to their Hebrew translations, i.e., (L1 → L2 translation) b. (Passive Recall) translating the target words into Hebrew, i.e., (L2 → L1 translation)	I. Immediate Post-test: a. Active Recall: Single words =4.31 (Max.= 10) Collocations=5.81 (Max.=10) b. Passive Recall: Single words=5.81 (Max.= 10) Collocations= 8.58 (Max.= 10) II. Delayed Post-test: a. Active Recall: Single words =4.12 (Max.= 10) Collocations=6.12 (Max.= 10) b. Passive Recall: Single words=6.27 (Max.= 10) Collocations= 8.73 (Max.= 10)	Compatible

Table 1: Summary of the initial study

A closer look at the above Table shows that in the initial study, the learning tasks and the retention tasks in the FFI and MFI groups were not compatible; in contrast, there was close compatibility between the learning tasks and the retention tasks in the CAT group. This phenomenon, as mentioned before, could have affected the results of the study, particularly those of the CAT group, whose performance was remarkably higher than the other two groups in the study.

Various comparative studies have focused on the relative efficiency of the Focus on Form conditions vis-à-vis Meaning Focused conditions. Their findings generally testify to the better retention results from the Focus on Form conditions (See Schmitt, 2008). However, the initial study's results, which demonstrated better performance of the CAT group over its counterpart, appear to be surprisingly high compared to other related studies. The mean retention rate in the CAT group was almost twice as much as the FFI group in the initial study. Because the FFI and the CAT groups are from the Focus on Form condition category, the initial study's results seem to be inflated, possibly due to the research design. These surprising findings motivated the author of the current study to replicate the initial study since, as Marsden et al. (2018, p.329) put it, "another factor potentially indicating importance, and thus a need for replication, are 'surprising' findings".

Many researchers believe that replication studies play an essential role in any scientific inquiry (Marsden et al., 2018). Some critics have depicted replication as the "gold standard" of research evidence (Jasny et al., 2011, p. 1225). Likewise, replication studies are deemed very scant, "with a mean rate of one published replication study for every 400 articles" (Marsden et al., 2018, p. 322). The success of the field of second language learning to come to vigorous and generalizable findings depends crucially on the validity and reliability of its research (Mackey and Gass, 2016). Additionally, there are two general types of replications: "direct replication," where no significant or intentional changes are made to the initial study, and "conceptual replication," where intentional alterations are made to the initial study to examine the generalizability of the findings to different conditions, contexts, and research features (Marsden et al., 2018). Thus, in response to the frequent calls from scholars in the field of second language acquisition to do replicational studies, and because of the methodological issues in Laufer and Girsai's (2008) study, it is necessary to design similar research to replicate the initial study. Because this study is conceived as a conceptual replication, the method of the initial study has been preserved. However, some key variables have been modified, specifically the format of the measurement tools (an utmost attempt was made to deny the test-task compatibility advantage from any of the groups), participants (a larger sample, more homogeneous in terms of language proficiency), and the target words (the number of the target single words was increased to sixteen). The research questions are as follows:

1. *Which vocabulary teaching method, i.e., MFI, FFP, or CAT will result in better retention of second language lexical items on an immediate post-test?*
2. *Which vocabulary teaching method, i.e., MFI, FFP), or CAT will result in better retention of L2 lexical items on a delayed post-test?*

Based on the initial study's findings, the expectation is that the CAT group would outperform the other two groups under all conditions, and the FF group would be the second-best.

Methodology

Participants

One-hundred five students from three intact (already-formed) groups participated in the study. Thirty six students were assigned to the MFI group, 34 to the FFI group, and 35 to the CAT group. They were both male and female, Persian native speakers, ages 18-22, studying English as a foreign language at universities or private language teaching schools in Tehran, Iran. The Structure and Written Expression, plus the Reading Comprehension Sections of a TOEFL test, were used to homogenize the participants in terms of language proficiency. The mean score was 70 (out of a maximum of 90), SD = 8.7. The participants whose scores fell within one standard deviation above or below the mean were assigned to the study, and others were deleted as outliers. These scores are comparable to the B2–C1 range of proficiency in the Common European Framework of Reference. Compared to the sample in the initial study (N= 75), the current participants were more proficient and, on average, four years older.

Materials

Pre-test and Target Items

As in the initial study, to ensure that the target words were relatively unknown to the participants, a pre-test consisting of 50 words was developed and administered to the groups two weeks before the treatment phase during their regular class time. Some words extracted from their course books also added to ensure that the students started thinking about the items; others functioned as distractors. The distractors were selected from among words the learners were supposed to be familiar with as they had been previously taught to them. In the initial study, the students had to write the L1 translation of the 50 English words in the pre-test, whereas in the replication study, the learners were required to provide as many L1 translations, or English synonyms of the words they knew in the word pool of their own volition. The pre-test took about ten minutes. Based on the pre-test, sixteen words were selected that were unknown to all of the participants.

The target words in the replication study were: *decapitate, frugal, cursory, brawl, interloper, liquidate, sedentary, inflame, dexterous, indolent, collaborate, distinct, antithesis, repulse, emerge, prolific.*

In the initial study, the ten target single words were as follows: *relish, glean, candid, laudable, opulent, plague, account, detractor, gregarious, lavish*.

In the replication study, the number of the target single words was increased to sixteen items to obtain more dependable statistical results.

Reading Text

Following the initial study, the sixteen target words were embedded in a text compiled by the author of the replication study. It was entitled *The Bees*. Complicated grammatical structures were avoided in the text, so that a lack of grammatical knowledge would not impede comprehension. The text contained 305 words, and enough textual clues enabled the readers to grasp the meaning without too much effort.

Procedure

An attempt was made to follow the same design, data collection procedures, and even statistical analysis as in the initial study by Laufer and Girsai. This study used a factorial design which Ary et al. (2013) define as "one in which the researcher manipulates two or more variables simultaneously in order to study the independent effect of each variable on the dependent variable, as well as the effects caused by the interactions among the several variables" (p. 311). The present research had one independent variable, i.e., the 'instruction type,' with three levels: MFI, FFI, and CAT. Also, the study embraced an incidental acquisition design, i.e., whether the target words were learned without participants' deliberate effort to commit them to memory. Also, the study embraced an incidental acquisition design, i.e., we whether the target words were learned without participants' deliberate effort to commit them to memory. Considering that the results of the study would be affected if the participants were told they were participating in an experiment, the requirement for informed consent was waived. It was decided that the experiment posed no risk to the participants, and it did not affect their rights and welfare in any way.

The treatment session was held in the regular class two weeks after the pre-test. The time-on-task (i.e., the amount of time allocated to each task) of treatment was constant for all groups.

The first stage of the treatment was similar across the three groups. The teacher briefly presented the topic and distributed the papers among the students. The papers contained the 305-word reading text titled The Bees followed by ten multiple-choice questions the participants had to answer (Task A). The participants were asked not to use dictionaries or other resources. However, if the students asked questions about the new words' meaning, the teacher provided sufficient information but did not elaborate on the single vocabulary. When the participants completed the exercise, the teacher went through all the questions and discussed the correct answers by referring students to the line or paragraph in which the question could be answered. This activity took about 40 minutes of the class time. All the teaching materials were collected back at the end of each session.

As in the initial study, the second stage of the study, which took 90 minutes, started the following day. All the groups received a different treatment consonant with the assigned experimental condition. The treatment was comprised of two tasks. The first task was completed with the reading text, to which the participants were exposed again. The second took place without the text. Similar to the first stage, the learners were not allowed to use dictionaries, and no marginal glosses for the target words were provided. If completing a task was deemed compulsory by the participants, they could guess the meanings from context or ask the teacher to clarify the words' meaning. As the teacher monitored all the participants' answers, the meaning of the target words was confirmed.

Following the initial study, seven different tasks were developed, the first of which (Task A) was common to all three groups. Task A consisted of the reading comprehension passage (*The Bees*) and ten true/false comprehension questions. Table 2 shows the tasks done by the students in the treatment sessions.

Groups	1st Task	2nd Task	3rd Task
Group 1 (MFI)	Task A	Task M1	Task M2
Group 2 (FFI)	Task A	Task F1	Task F2
Group 3 (CAT)	Task A	Task C2	Task C2

MFI = Meaning Focused Instruction; FFI = Form Focused Instruction; CAT = Contrastive Analysis and Translation

Table 2: Tasks in the treatment session

For our MFI group, in addition to Task A, which consisted of the reading comprehension passage *The Bees* and ten true/false comprehension questions common to all three groups, two extra tasks were designed to be implemented as the MFI activities. First, M1 was a reading comprehension task in which the students were supposed to produce answers to eight questions according to the contents of reading passage. Below are two examples of the questions in M1:

1. In paragraph 2, why does the writer say: "It's a cruel, cruel world"?

2. Complete the following sentence:

When a female wants to come out of the soil _____.

As seen in the above examples, to answer the questions, the students might have needed to think of the new words embedded in the text, but the questions did not draw the students' attention to the single target words in the text.

The M2 task designed for our MFI group was a pair/group communicative task in which the students had to discuss some controversial issues. An example of the topic is as follows:

3. In your opinion, which of the two bee colonies is more similar to human societies, the honeybee or Dawson's bee?

For the FFI group, two form-focused tasks were designed: the first one (F1) was a task on recognizing the target words in the multiple-choice format. The following example is taken from F1:

4. ***Dexterous****:*

a. skillful at using hands b. able to build new things C. good at problem-solving

In order to do this task, the students in the FFI group had to consider the contextual clues in the text and choose the best answer out of the three given choices.

The second task (F2) for the FFI group was a fill-in-the-blank task in which the students had to choose the answers from a 'word bank' given in a box above the task. For instance, the word 'inflame' had to be used in the following example:

5. The president tried to _____ public opinions about the new law.

The sentences in this task were close to the ones the students had seen in the text, and they had to figure out which word best suited the given blank.

For the third group (CAT), two translation tasks were developed: The first task (C1) was a series of sentences in English containing the target words. The students were supposed to provide an L1 translation of each sentence. The target item is in bold font.

6. The pianist was really **dexterous**. _____

The second task (C2) asked the students to translate L1 sentences into L2. Again, the translations of the target words were made bold in the L1 sentences, and the students were told to use the target words in their L2 corresponding sentences.

One of the challenges faced in the replication study was to design a test in which no advantages were granted to any one of the groups in the study. Thus, half of the items were designed in the translation format, just as in the initial study, and the other half resembled the non-contrastive form-focused tasks, such as fill-in-the-blank and multiple-choice questions. The test, which was developed to measure the immediate and delayed retention of the target items, had two sections: the first section measured the active recall of the words by the students, and the second section measured the passive recall. Although the items that comprised the immediate and delayed retention post-tests were the same, the order of the items was reshuffled in the delayed post-test to reduce the possible memory effect which might have accrued as a result of the participants taking the same test twice.

The active recall of the students was measured through two tests: A translation test which exposed the students to a list of Farsi words and asked for the English equivalents (exactly in the same format as the initial study), and a blank-filling test which was developed in the replication study. It contained some sentences with a blank in which one of the target words was required. Here are two examples of the active recall test:

Translate the following Farsi words into English:

صرفه جو *(sarfe-joo): _____*

Fill in the blanks with the most appropriate words you encountered in the passage and the exercises.

The police caught the _____who had gone into the hospital illegally.

The passive recall of the students on the target words was measured through a test in two formats: one in which the students had to choose the best meaning inferred from a sentence containing the target item in the multiple-choice format--introduced in the replication study--and the second format in which the students were required to translate the target words from English into Farsi. The second format was similar to the one used in the initial study.

Here are two examples of the passive recall test:

Choose the best answer that matches the meaning of the underlined word.

*They managed to **repulse** the enemy's attack.*

a. They recognized the attack b. They stopped the attack c. They carried out the attack

Translate the following English words into Farsi:

Decapitate _____

The passive and active recall tests were administered as the immediate and delayed post-tests to measure the students' immediate/delayed active and immediate/delayed passive recall of the target items. The testing procedures took almost twenty minutes for each phase (immediate and delayed) in all the classes. Following the initial study, the delayed post-test was administered one week after the immediate recall test.

As in the initial study, the answers were scored dichotomously, i.e., one point was assigned to a correct answer and zero points to an incorrect one; the items with no answers were treated as incorrect. In the active recall test, minor spelling errors that did not significantly alter the target word's pronunciation were neglected. For example, the answer s*edentery* for 'sedentary' was given one point, whereas *repalse* for 'repulse' was not. Since the main objective of the test was to tap the knowledge of the meaning-form link, and those spelling errors did not alter the word's form considerably, the answers were assigned one point. Non-target synonyms for the Farsi prompts were assigned zero points, for instances 'appear' instead of 'emerge.' Thus, the participants were told in advance of the test to use words only from the text they had studied, and they did so accordingly. In the passive recall test, any Farsi translation which indicated the semantic content of the intended target word was assigned one point. The highest score for the test was sixteen, and the internal consistency reliability index calculated using Cronbach's alpha turned out to be 0.86.

Results

Descriptive Statistics

In harmony with the initial study, the descriptive statistics of the data from the replication study, as shown in Table 3 and Table 4 below, indicated that the highest mean in the immediate post-test was that of the CAT group. The lowest mean came from the MFI group on both the passive and active recall tests. The CAT group had the highest mean on the immediate test compared to the other groups. However, unlike in the initial study, the mean difference between the CAT and FFI groups was not considerably large in the replication study. Like the initial study, the lowest mean was found in the MFI group, and both the FFI and the CAT groups scored better results in the retention of the target items than the MFI.

	Replication Study					**Initial Study**		
	Max = 16					Max = 10		
	N	**M**	**(SD)**	**(%)**	**N**	**M**	**(SD)**	**(%)**
MFI	36	3.31	(2.08)	(20%)	26	0.31	(0.68)	(3%)
FFI	34	11.03	(4.04)	(70%)	23	4.04	(2.64)	(40%)
CAT	35	11.37	(3.5)	(71%)	26	5.81	(2.48)	(58%)

Note. N = Number of participants; M = Mean; SD = Standard Deviation; % = Percentage of the retained items; MFI = Meaning Focused Instruction; FFI = Form Focused Instruction; CAT = Contrastive Analysis and Translation

Table 3: Descriptive statistics of the immediate post-test scores of passive recalls in the replication and initial studies

	Replication Study					**Initial Study**		
	Max = 16					Max = 10		
	N	**M**	**(SD)**	**(%)**	**N**	**M**	**(SD)**	**(%)**
MFI	36	0.94	(0.95)	(6%)	26	0.12	(0.33)	(1.2%)
FFI	34	9.47	(4.0)	(59%)	23	2.30	(2.18)	(23%)
CAT	35	10.40	(3.5)	(65%)	26	4.31	(2.51)	(43%)

Note. N = Number of participants; M = Mean; SD = Standard Deviation; % = Percentage of the retained items; MFI = Meaning Focused Instruction; FFI = Form Focused Instruction; CAT = Contrastive Analysis and Translation

Table 4: Descriptive statistics of the immediate post-test scores of the active recalls in the replication and initial studies

	Replication Study					**Initial Study**		
	Max = 16					Max = 10		
	N	**M**	**(SD)**	**(%)**	**N**	**M**	**(SD)**	**(%)**
MFI	36	2.31	(1.26)	(14%)	26	0.15	(0.37)	(1.5%)
FFI	34	11.91	(2.96)	(74%)	23	3.52	(2.94)	(35%)
CAT	35	12.34	(3.21)	(77%)	26	6027	(2.89)	(63%)

Note. N = Number of participants; M = Mean; SD = Standard Deviation; % = Percentage of the retained items; MFI = Meaning Focused Instruction; FFI = Form Focused Instruction; CAT = Contrastive Analysis and Translation

Table 5: Descriptive statistics of the delayed post-test scores of the passive recalls in the replication and initial studies

Scrutiny of the results from the initial research by Laufer and Girsai (2008) showed that the ratios of the retained target words (shown in percent in the brackets) were generally lower than the corresponding ratios obtained from the replication study. This was probably rooted in the different post-test formats employed in the two studies. In the initial study, the participants were asked to provide translations (or L2 definitions of the participant's own volition) for the target words, a testing format which is productive and open-ended. However, in the replication study, I used a combination of formats such as translation into /or from L2, contextualized fill-in-the-blank, or multiple-choice questions, almost half of which were of the recognition type. Recognition tests are believed to offer better chances to the test test-takers in to identify the correct answers from among the distractors than in the productive type tests. According to Clariana and Lee (2001, p. 24), "… recognition posttest scores of specific memory content will be greater and often far greater than recall [production] scores" (See also Heidari-Shahreza and Tavakoli, 2016; Martinez and Katz, 1996). For this reason, the ratio of the retained words in the current research was probably higher than those in the initial study.

Also, as is evident from the descriptive data in the above tables, the means of active recall scores in both the immediate and delayed post-tests were lower than those of passive recall scores across the three groups. This was not unexpected because, as Laufer and Girsai (2008) contended, "Vocabulary learning is an incremental process, and learners usually acquire passive knowledge of a word before they acquire its active knowledge (except in the case of cognates)" (p. 707).

Interestingly, in line with the initial study's findings, the participants in both the FFI and the CAT groups scored higher in the delayed post-test than in the immediate post-test of passive recall. Because all the materials were taken away after each session of the study, a possible justification for the higher scores in the delayed post-test might be the heightened motivation of the students to learn the target items they could remember from the treatments in other contexts.

Comparison of Means: ANOVA Results

In line with the initial study, to answer the first and second research questions, four one-way ANOVAs were run, two for the immediate scores and two for the delayed scores, to examine the mean differences between the three groups. However, before running the ANOVA procedure, the data were checked through Kolmogorov-Smirnov's Test for the normality of distribution and Levene's Test for the homogeneity of variances. The results indicated that the data were normally distributed, and the variances were homogeneous.

ANOVAs were run for the following data: Immediate-Active Recall, Immediate-Passive Recall, Delayed-Passive Recall, and Delayed Active Recall. As shown in Table 6, all the ANOVAs showed a significant difference between the means of scores across the teaching conditions.

	Passive Recall	**Active Recall**
	F value - (η_p^2)	**F value - (η_p^2)**
Immediate	233.38* - 0.82	100.57* - 0.66
Delayed	166.52* - 0.76	43.49* - 0.46

Note. * $p < 0.00$

Table 6: Differences between the three conditions: ANOVA results

Table 6 shows the F values for the ANOVAs on the effect of the teaching conditions on the retention of scores. A significant effect of instruction condition was observed on the passive recall of the lexical items in the immediate post-test [F (2,102) =233.38, p= 0.00], (η_p^2) = 0.82. The partial eta squared (η_p^2) revealed that 82% of the variance in the data on the immediate test scores in the passive recall was accounted for by the teaching conditions.

There was also a significant effect of instruction condition on the immediate active recall of the items [F (2,102) =100.57, p= 0.00], (η_p^2) = 0.66. The effect of instruction condition on the delayed test on the passive recall of the vocabulary items was also significant [F (2,102) =166.52, p= 0.00], (η_p^2) = 0.76. A significant effect of instruction condition was also observed on the delayed active recall of the lexical retention [F (2,102) =43.49, p= 0.00], (η_p^2) = 0.46. Because the F values for the ANOVAs were significant

($p < 0.00$), the Tukey was run post hoc to examine the mean differences between the group pairs. (See Tables 7 and 8).

		Replication Study		Initial Study	
		Passive Recall	Active Recall	Passive Recall	Active Recall
Group I	Group J	MD (I-J)	MD (I-J)	MD (I-J)	MD (I-J)
MFI	FFI	-9.665*	-8.526*	-3.74***	-2.19***
	CAT	-10.037*	-9.456*	-5.50***	-4.19***
FFI	MFI	9.665*	8.526*	3.74***	2.19***
	CAT	-.372	-.929	-1.76**	-2.00***
CAT	MFI	10.037*	9.456*	5.50***	4.19***
	FFI	.372	.929	1.76**	2.00***

Note. * $p < 0.001$; MD = Mean Difference, Std. Error = 0.6

Table 7: Mean difference between pairs of instruction conditions on the immediate post-test in the replication and initial Study

As is evident from Table 7 above, the replication study results from the immediate post-test on the passive and active recall showed the differences between means of the MFI and FFI groups, and the means difference between the MFI and CAT groups, were significant at $p < 0.001$. These findings confirm the results of the initial study on the corresponding comparisons. In line with the initial study, the highest difference of means was found in the MFI and the CAT pairs. However, unlike in the initial study, there was no significant difference between the means of FFI and CAT groups on the immediate post-test, nor for the active or passive recall of the items.

Also, on the delayed post-test of the active recall, the mean differences of the FFI and CAT groups turned out to be non-significant in contrast to the initial study's findings. However, in harmony with the initial study, the differences between all pairs of conditions were significant in the delayed passive recall in the replication study.

		Replication Study		Initial Study	
		Passive Recall	Active Recall	Passive Recall	Active Recall
Group I	Group J	MD (I-J)	MD (I-J)	MD (I-J)	MD (I-J)
MFI	FFI	-9.606*	-3.474*	-3.37***	-2.64***
	CAT	-10.094*	-6.273*	-6.12***	-5.54***
FFI	MFI	9.606*	3.474*	3.37***	2.64***
	CAT	-.488	-2.779*	-2.75**	-2.89***
CAT	MFI	10.094*	6.273*	6.12***	5.54***
	FFI	.488	2.779*	2.75***	2.89***

Note. * $p < 0.001$; MD = Mean Difference, Std. Error = 0.6

Table 8: Mean difference between pairs of instruction conditions on the delayed post-test in the replication study and the initial study

Hence, the results of the inferential statistics from the replication study were only partially consistent with those of the initial study, whose findings are presented in Tables 7 and 8 above. The CAT and the FFI method turned out to be the most effective, a finding in line with the initial study. However, the mean difference between the FFI and CAT pairs, which appeared statistically significant in all conditions in the initial study, proved to be significant only on the delayed active recall in the replication research. To put it into different words, the FFI and the CAT methods yielded roughly equal outcomes except for the delayed active recall of words. The apparent inconsistency can be attributed to the different measurement methods adopted in the two studies. In the initial study, there was close compatibility between the tasks in the treatment and the test method used to measure retention. In other words, the participants who did translation tasks and received explanations on the aspects of contrastive analysis from the teacher answered translation questions in the exam. Hence, they took advantage of the awareness of contrastive analysis gained from the instruction sessions on the post-tests. This advantage probably culminated in the better performance of the CAT group than the FFI one. In the replication research, however, an attempt was made to deny the task-test compatibility privilege from the CAT group.

Consequently, the means of the two groups at issue became almost equal, except for the delayed active recall. In sum, the overall findings from these two studies indicate that the CAT method is as effective as

the FFI (and sometimes even more effective) in teaching/learning vocabulary and that it might lead to more durable learning outcomes. This claim can be justified by the Lexicalization Hypothesis proposed by Paribakht (2005). Lexicalization can potentiate the target word's associative relations to the existing L1 lema (with the proviso that there is a close concordance between the semantic patterns of the target L2 word and those of the L1), which in turn can increase the likelihood that the target L2 word be retained more successfully in later retrievals (see Heidari-Shahreza & Tavakoli, 2016; Paribakht, 2005).

Discussion

In response to the first research question, *Which vocabulary teaching method, i.e., MMF, FFI or CAT, will result in a better retention of second language lexical items on an immediate post-test?*, the CAT and the FFI proved to result in roughly equal means which were hugely superior to MFI in the active and passive recalls. The rank order of the three methods in terms of efficiency was: FFI = CAT > MFI. This is partly inconsistent with the initial study's findings in which a large magnitude of the difference was found between the means from the FFI and CAT groups on the immediate active and passive recalls. The rank order of the methods in the initial study was as follows: CAT > FFI > MFI.

To address the second research question: *Which vocabulary teaching method, i.e., MMF, FFI or CAT, will result in a better retention of second language lexical items on the delayed post-test?*, the data from the delayed post-test in the replications research revealed the same rank order as the immediate post-test on the passive recall, i.e., the CAT and the FFI equally yielded the largest means and the MFI the lowest. The rank order of the methods from the viewpoint of efficiency on the delayed passive recall was in the following order: CAT = FFI > MFI. However, on the active recall, the rank order of the methods in terms of efficiency changed in the following manner: CAT>FFI>MFI. The CAT yielded the highest mean; the FFI means ranked second, and the MFI resulted in the lowest mean. The rank order of the three methods in the initial study was: CAT >FFI >MFI across active and passive recalls.

Additionally, since the initial study had not taken the test-task compatibility effect into account in designing the post-tests, it was likely that the CAT group, whose task included translation and contrastive analysis, outperformed the other groups not because the CAT method was more effective, rather because they were required to do the same things on the post-test as they had practiced in the tasks in the treatment sessions. Hence, a post-test was designed with a mixture of test formats used so that it would not grant any specific privilege to any of the groups due to the test-task compatibility effect. Consequently, in the replication research, the effect of the CAT and the FFI on word retention was almost the same. No significant advantage for the CAT method was found except for the delayed active recall of words. The implication is that test-task compatibility is an important factor that can change the measurement outcomes in research studies and might lead to different interpretations of the findings.

Several theoretical bases support the findings of the replication research and the initial study. The first justification would be the noticing hypothesis proposed by Schmidt (2001). This hypothesis tries to provide a theoretical framework for FFI and contends that learners need to consciously notice forms and the senses that these forms designate in the input to change input into the intake for learning (Schmidt, 1990). Schmidt (2001) defines 'noticing' as the subjective parallel of the term attention and the ranges of meanings it signifies in psychology and settles on the conclusion that although some kinds of learning might occur without attention, in a broad-spectrum, attention coupled with intention is probably a practical requisite for effective language acquisition. Because many L2 input features are sporadic, lack saliency, and are redundant in linguistic interactions, they might go overlooked unless the learner focuses on them. In the current study, the attention of the FFI and the CAT group participants was drawn to the forms of the target items by singling the words out in different tasks. However, no lexical items in the reading text were singled out in MFI. As a result, the lack of noticing decreased the efficacy of the MFI group in vocabulary acquisition. Before a language feature enters the long-term memory, it must be noticed, identified, and practiced in the short-term memory (Robinson, 1995). The performance of the MFI group in the immediate post-test showed that most failed to commit the lexical items to their related schemata. Another justification for the superiority of form-focused instruction is proposed by VanPatten (2002), who maintains that the restricted capacity of human processing memory does not allow sufficient attention to the form of the linguistic items when learners are performing a communicative task.

In the same vein, the apparent superiority of the translation can be explained within the framework of the 'pushed output' hypothesis. The 'pushed output' hypothesis states that when learners use language in communicative interactions and expand their linguistic repertoire in the process, they enhance their language production abilities and boost their language learning rate (Swain, 1985). James (2005) believes that translation is a perfect manifestation of pushed output, giving way to measuring the active recall of the linguistic items. According to de la Fuente (2002), and Ellis and He (1999), there are amply documented findings suggesting that output tasks are more influential than input-oriented tasks in the retention and recall of new words. As Laufer and Girsai (2008) contended, "In order to translate, the learner is required to produce language, but unlike the case of free production, the learner cannot produce a good translation if he avoids problematic words or structures. Hence, translation should be at least as effective as other pushed output tasks for learning vocabulary" (p. 698).

The findings of the current research like those of the initial study regarding the superiority of form-focused instruction in promoting the retention of vocabulary items over the meaning-focused approaches are in line with a plethora of different studies (Li, 2010; Loewen & Philp, 2006; Long, 1991; Lyster & Saito, 2010; Mackey & Goo, 2007; Mackey & Philp, 1998; Valezy & Spada, 2006; Swain, 1988). Although most of the studies mentioned discuss form-focused instruction for grammatical or general linguistic forms rather than specifying the retention of vocabulary, the current study's findings have shown that vocabulary learning is no exception to the general trend. The FFI conditions can enhance it.

Conclusion

The influence of three different types of instruction (MFI, FFI, and CAT) was observed on the short- and long-term retention of newly presented vocabularies embedded in a reading text. The conclusion drawn from the present study is threefold. First, the comparison of different instructional methods revealed that contrastive analysis and translation was the most effective method of instruction in teaching vocabulary. Though ranked second in efficiency, FFI turned out to be far superior to meaning-focused instruction. Secondly, the results of the study remind us of the fact that the role of translation and contrastive analysis in language teaching, which is still a controversial issue and has been ostracized into the corner of negligence for many years (see Cook, 2010), needs to be reappraised and re-conceptualized in studies of second language acquisition. The findings of the current research, plus those of Laufer and Girsai (2008), imply that the role of translation in second language acquisition should not be ignored and that the importance of L1 in learning L2 is undeniable. Finally, it was shown in the study that compatibility between a task and the testing method might manipulate the results in the sense that in research in which the purpose is to measure the effectiveness of one method in comparison to another one, the choice of the very same type of methods practiced in the treatment for the later assessment of learning might lead to overestimated results due to the effect of the testing method. It was shown that the compatibility of the task and the testing method gives an undue advantage to one group of participants over the other(s).

It was also shown that focus on form—any method that draws students' attention to the features of language components such as part of speech, pronunciation, dictionary meaning, roots of the words, equivalents in the mother tongue, etc.—might be very helpful, at least, in vocabulary learning.

One of the limitations of the replication study is that intact classes were assigned to the research. Therefore, there was no control over the participants' gender, language aptitude, educational backgrounds, or the languages they knew other than English, or their mother tongue, which are all important factors influencing the learner's performance in vocabulary learning. Thus, further research with a more rigorous design might be necessary to refine the issue. There is also an opportunity to implement different testing methods with varying test-task compatibility degrees and to investigate how they can influence word retention outcomes. An additional suggestion in this respect could be using computational programming in which the response time to test items of different formats could be measured, and more exact results could be obtained. Furthermore, in future research, it would be interesting to investigate learners' attention to the target words, e.g., eye tracking, during reading when they are doing the FF, MF, or CAT tasks.

Acknowledgement

I would like to extend my thanks to Mr. Yaser Shamsi who helped a lot in teaching the classes in different task groups, preparing the materials, and gathering the data in this research study. I am also immensely grateful to JoAnn Miller (Editor-in-Chief) and Josefina C. Santana Villegas (Deputy Editor-in-Chief) for their insightful comments and suggestions to improve the paper.

References

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2013). *Introduction to research in education*. Cengage Learning.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331-354). Lawrence Erlbaum.

Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development, 49*(3), 23-36. https://doi.org/10.1007/BF0250491

Cook, G. (2010). *Translation in language teaching: An argument for reassessment*. Oxford University Press.

de La Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition*, *24*(1), 81-112. https://doi.org/10.1017/S0272263102001043

Ellis, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, *21*(02), 285-301. https://doi.org/10.1017/S0272263199002077

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

Heidari-Shahreza, M. A., & Tavakoli, M. (2016). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition by Iranian EFL Learners. *The Language Learning Journal, 44*(1), 17-32. https://doi.org/10.1080/09571736.2012.708051

Hemchua, S., & Schmitt, N. (2006). An analysis of lexical errors in the English compositions of Thai learners. *Prospect, 21*(3), 3–25.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 356-357). Blackwell.

James, C. (2005). Contrastive analysis and the language learner: A new lease of life? In D. J. Allerton, C. Tshichold, & J. Wieser (Eds.), *Linguistics, language learning and language teaching,* (pp. 1-20). Broschiert.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again . . . *Science*, *334*, 1225–1225. https://doi.org/10.1126/science.334.6060.1225

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

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694-716. https://doi.org/10.1093/applin/amn018

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta analysis. *Language Learning*, *60*(2), 309-365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Liao, P. (2006). EFL learners' beliefs about and strategy use of translation in English learning. *RELC Journal, 37*(2), 191–215. https://doi.org/10.1177%2F0033688206067428

Loewen, S., & Philp, J. (2006). Recasts in the adult English L2 classroom: Characteristics, explicitness, and effectiveness. *The Modern Language Journal*, *90*(4), 536-556. https://doi.org/10.1111/j.1540-4781.2006.00465.x

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R, B. Ginsberg, & C. Kramsch (Eds.). *Foreign language research in cross-cultural perspective*, (pp. 39-52). John Benjamins.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265-302. https://doi.org/10.1017/S0272263109990520

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). Routledge.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 407–452). Oxford University Press.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *The Modern Language Journal*, *82*(3), 338-356. https://doi.org/10.1111/j.1540-4781.1998.tb01211.x

Makel, M., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(1), 537–542. https://doi.org/10.1177%2F1745691612460688

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning, 68*(2), 321-391. https://doi.org/10.1111/lang.12286

Martinez, M. E., & Katz, I. R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment, 3*(1), 83–98. https://doi.org/10.1207/s15326977ea0301_4

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533. https://doi.org/10.1016/S0022-5371(77)80016-9

Paribakht, T. S. (2005). The influence of first language lexicalization on second language lexical inferencing: A study of Farsi-speaking learners of English as a foreign language. *Language Learning, 55*(4), 701-748. https://doi.org/10.1111/j.0023-8333.2005.00321.x

Puig-Mayenco, E., González Alonso, J., & Rothman, J. (2018). A systematic review of transfer studies in third language acquisition. *Second Language Research, 36*(1), 31-64. https://doi.org/10.1177%2F0267658318809147

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161. https://doi.org/10.1017/S0267190504000078

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*(2), 283-331. https://doi.org/10.1111/j.1467-1770.1995.tb00441.x

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge Applied Linguistics.

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

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

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Newbury House.

Swain, M. (1988). Manipulating and complementing content teaching to maximize second language learning. *TESL Canada Journal*, *6*(1), 68-83. https://doi.org/10.18806/tesl.v6i1.542

Valezy, J. R., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). John Benjamins.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning*, *52*(4), 755-803. https://doi.org/10.1111/1467-9922.00203

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, 52, 203–226. https://doi.org/10.1002/rrq.163