Task-based incidental vocabulary learning in L2 Arabic: The role of proficiency and task performance

Ayman A. Mohamed
Michigan State University

Abstract

This study tests the claim that word learning in a second language are contingent upon a task’s involvement load (i.e. the amount of need, search, and evaluation it imposes), as proposed by Laufer and Hulstijn (2001). Fifty-three English-speaking learners of Arabic were assigned to one of three vocabulary learning tasks that varied in the degree of involvement: reading comprehension with glosses (low), fill-in-the-gap task (medium), and sentence writing (high). Ten words, selected based on a pretest, were targeted in the tasks. Results showed a main effect of task, with the sentence writing task yielding the highest rates of vocabulary learning, followed by the gap-fill task, and finally the reading comprehension task. A significant correlation was found between accuracy of performance across participants and their subsequent vocabulary acquisition in the immediate posttest. Within groups, only the performance of the writing group correlated significantly with their posttest scores. Results of the present study validate the hypothesis and point to multiple factors at play in incidental vocabulary acquisition. The study provides further arguments to refine the hypothesis and implement pedagogical practices that accommodate incidental learning in foreign language settings.

Keywords: incidental learning, L2 Arabic, vocabulary acquisition, involvement load
Task-based incidental vocabulary learning in L2 Arabic: The role of proficiency and task performance

From a pedagogical point of view, vocabulary studies reflected the need to identify which teaching tasks provide better opportunities for vocabulary learning. The majority of research on incidental vocabulary learning investigated various ways to manipulate input through tasks with hypothesized differential effects (Cho & Krashen, 1994; Hulstijn, Hollander, & Greidanus, 1996; Hulstijn & Tromper, 1998; Peters, Hulstijn, Sercu, & Lutjeharms, 2009). Results have shown that tasks inducing more effort on the part of learners yielded more significant learning outcomes. Schmitt (2008) interpreted this factor in terms of engagement with vocabulary in learning tasks. In a wider framework, Laufer and Hulstijn (2001) introduced the involvement load hypothesis as a motivational cognitive construct that can predict the amount of incidental intake from text-based tasks. Several lines of empirical evidence were presented (Hulstijn & Laufer, 2001; Kim, 2008; Keating, 2008) to validate the hypothesis. However, other studies refuted the involvement load as an overriding factor in vocabulary learning, pointing that the type of task (input-based vs. output-based) must be considered in calculating task-induced involvement (Folse, 2006; Yaqubi, Rayati, & Gorgi, 2011).

The purpose of the current study is to investigate incidental vocabulary acquisition in the light of the involvement load hypothesis while factoring in the role of the learner-related variations in task performance and proficiency levels, which have not been considered in previous studies. An important contribution of the study is the inclusion of learners of Arabic as a less commonly represented population in second language research.
Incidental vocabulary learning

Incidental vocabulary acquisition is defined by Paribakht and Wesche (1999) as the potential of acquiring new lexical items as a byproduct of focusing on understanding meaning rather than on the explicit goal of learning new words. There has been a widely accepted view that, apart from high frequency vocabulary, the core of lexical development occurs incidentally, particularly through reading (e.g., Fraser, 1999; Hill & Laufer, 2003; Horst, 2005; Huckin & Coady, 1999; Kweon & Kim, 2008; Matsouka & Hirsh, 2010; Nation, 2001; Paribakht & Wesche, 1999; Pulido, 2007; Rott, 1999; Pellicer-Sánchez & Schmitt, 2010; Watanabe, 1997; Webb 2008; Zimmerman, 1997). Although incidental learning remained controversial and less efficient in many ways (Ellis, 1999; Gass, 1999; Laufer, 2005; Macaro, 2003; Read, 2004), it has become pedagogically acknowledgeable that both modes of learning complement each other in the process of learners’ incremental vocabulary development (Schmitt, 2008, 2010).

Task-based incidental vocabulary learning

Previous vocabulary studies investigated the role of different language modalities in providing opportunities for incidental learning. In oral input, it was found that incidental vocabulary acquisition occurred as a byproduct of negotiation and output within interaction and speaking tasks (Ellis, Tanaka, & Yamazaki, 1994; Ellis & He, 1999; de la Fuente, 2002; Brown, Sagers, & LaPorte, 1999). Listening tasks were found to be conducive to vocabulary learning yet with lower rates than interaction tasks (Brown, Waring, & Donkaebua, 2008; Elley, 1989; Smidt & Hegelheimer, 2004; Vidal, 2010). Some classroom research reported learning outcomes from spontaneous class interaction and teaching activities (e.g., Dobinson, 2001; Mohamed, 2012). Horst (2010) supported this line of research through a corpus-based study that indicated many opportunities for incidental intake from teacher-talk and classroom communication.

Text-based tasks were more frequently investigated in vocabulary studies. Research in this area focused on how to promote
engagement in reading tasks, either by manipulating word presentation and saliency in texts or administering different tasks with varying degrees of complexity. For example, learners who inferred the meanings of certain words by having to choose from options provided retained words better than the other group who were only provided the meanings of target words in a gloss (Hulstijn, 1992). Looking up meanings in a dictionary was a more effective task than encountering meaning in marginal glosses (Hulstijn, Hollander, & Greidanus, 1996). Reading followed by vocabulary-focused exercises yielded better retention than reading with inferring meaning from context (Paribakht & Wesche, 1997). Reading combined with dictionary usage was more beneficial than reading only (Cho & Krashen, 1994; Knight, 1994; Lupescu & Day, 1993). Using words in a composition was more effective than only encountering words in reading comprehension (Hulstijn, & Trompetter, 1998). To find a general interpretation of the common findings in vocabulary studies, Schmitt (2008) referred to engagement with lexical items as a key factor in vocabulary learning. Engagement, in his view, can be fostered by many factors, including, but not confined to, frequency of exposure, increased attention to target words, and increased time spent on the target items. Watanabe (1997) and Peters, Hulstijn, Sercu and Lutjeharms (2009) found that the text input which affords increased processing due to contextual, lexical, or semantic enhancement is more likely to yield more vocabulary gains (see Rott, Williams, & Cameron, 2002; Rott & Williams, 2003).

Taken together, Schmitt (2008) maintains that the common factor in the results of these studies can be termed as engagement with vocabulary. He mentions some forms of engagement suggested by empirical studies. The forms include increased frequency of exposure, increased attention focused on the target items, increased noticing of the lexical items, increased intention to learn lexical items, need to learn the lexical item, increased manipulation of word properties, increased amount of time spent with the lexical item, and the amount of interaction spent on the lexical item. The more learners are engaged with new words, the more likely the words will be acquired and retained.
Involvement Load Hypothesis

Beyond paper-and-pencil results, some research was more interested in cognitive interpretations for vocabulary learning outcomes. Laufer and Hulstijn (2001) attempted to interpret the findings of previous empirical studies in a systematic manner by dividing tasks into more effective and less effective tasks and exploring the common characteristics of specific tasks. They eventually introduced the Involvement Load Hypothesis to account for the role of involvement in learning tasks that establish conditions for incidental learning. The Involvement Load Hypothesis claims incidental vocabulary acquisition can only take place when there is cognitive processing—that is, One’s processing of the word must be involved in some way. In this sense, Laufer and Hulstijn (2001) noted that studies on vocabulary acquisition compared tasks with higher processing requirements and tasks with lower processing loads. Although these studies did not directly test the hypothesis, their results, according to Laufer and Hulstijn (2001), are in favor of the tasks that allow learners to be more involved with the target words while completing the activities. They argued that acquisition is a function of the degree of involvement rather than the nature of the task (i.e., either input-oriented or output-oriented); involvement being an overriding factor in vocabulary learning.

In an effort to operationalize the proposed motivational-cognitive construct, Laufer and Hulstijn divided involvement load into three components: Need, Search, and Evaluation. Need is a motivational component that implies the need for a specific word to complete a certain task. Need is moderate (1) if it is externally imposed (e.g., the teacher asks the students to fill in the word in a sentence). Need is strong (2) when it is internally imposed (e.g., the learner decides to look up the word and use it in a given task). Search refers to trying to find information like looking a word up in a dictionary or asking the teacher. It is a cognitive component that is described as absent (0) or present (1), and was not hypothesized to have degrees. Evaluation refers to comparing a word to its context, to other words, or to its other meanings and determining if it fits. It is described as moderate (1) if it requires the learner to compare
between words, and it is described as strong (2) if it entails making decisions about how the word fits in combination with other words in a given context.

To measure the involvement load, an involvement index is calculated by summing up the amount of need, search, or evaluation presented by a given task. Each component is assessed as absent (0), moderate (1), or strong (2). According to the hypothesis, tasks with a higher involvement index are predicted to yield more word acquisition and better retention than lower-index tasks.

**Empirical evidence for the Involvement Load Hypothesis**

Laufer and Hulstijn (2001) provided an empirical model through two experiments with advanced EFL learners performing three different tasks: (i) reading comprehension with glosses (moderate need, no search, and no evaluation); (ii) reading plus filling in target words (moderate need, no search, and moderate evaluation); and (iii) composition writing using target words (moderate need, no search, and strong evaluation). Their results supported the hypothesis, as the amount of retention was related to the total involvement load. The group that worked on composition retained the most, followed by the reading-plus-gap-fill group, and finally the reading-with-gloss group. Kim’s (2008) replication supported the hypothesis, adding that the effectiveness of different tasks for vocabulary learning was not affected by learner proficiency. Keating (2008) similarly corroborated the role of the involvement load and found that beginners could also benefit from demanding tasks to improve their vocabulary.

On the other hand, Folse (2006) found counterevidence to the predictions of the Involvement Load Hypothesis, in that was as effective as using new words to write original sentences (strong evaluation) supplying new words in a fill-in-the-gap task (moderate evaluation). In particular, he found that completing three fill-in-the-gap activities yielded more gains than completing one sentence-
writing activity. Folse thus argued that word learning and retention has more to do with frequency of exposure and word recycling in various activities rather than the involvement load of the tasks themselves. Yaqubi, Rayati and Gorgi (2010) suggest that the involvement load may not be an overriding effect in vocabulary learning. They found variations in the effectiveness of tasks due to their type, whether they were input or output-based, rather than their involvement loads.

The above studies suggest that the involvement load hypothesis needs empirical research to test its components as they interact with other factors. One issue that has not been fully investigated through the aforementioned studies is the effect of proficiency level on the ability to benefit from high-involvement tasks. Moreover, no study in this framework has referred to the relationship between learners’ performance on the tasks themselves and the predicted vocabulary outcome of the task.

The Involvement Load Hypothesis was originally informed by principles from the depth of processing hypothesis (Craik and Lockhart, 1972) which distinguished shallow and deep processing in task performance and proposed that semantic processing is associated with higher levels of retention for target items. Early empirical investigation of the hypothesis by Craik and Tulving (1975) indicated that learner’s positive responses to questions on target words was associated with higher levels of retention than negative responses. Barcroft (2002) maintained that increased semantic processing in task performance may inhibit encoding of the formal properties of target words. A common line that can be drawn from these studies that task performance does affect the learning outcome demonstrated by learners as they develop their linguistic resources. I hypothesize that the accuracy of performance that learners show in the learning tasks would positively reflect on the number of words retained as measured by vocabulary knowledge tests.
Current study

The goal of this study is to investigate incidental vocabulary acquisition in the light of the involvement load hypothesis while exploring the additional factors of task performance and learners’ proficiency levels. In this sense, the study is testing whether the involvement load is an overriding factor in vocabulary acquisition while extending the research to new contexts in vocabulary instruction. The current curricula in Arabic language instruction suffer from the unsystematic presentation of vocabulary and a considerable lack in task-based material. One of the goals of the study is to pilot the adaptability of given learning tasks to the Arabic teaching material and investigate what learners need most in terms of enhancing their lexical knowledge in a foreign language setting.

The experimental design will be set to accommodate a pretest-task-posttest sequence. Table 1 details the design of the study along with descriptions of tests used and timeline of each procedure. The details of material and procedures are presented in the following sections.

Table 1: Experimental design and procedure timeline

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Procedure</th>
<th>Description</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two weeks before the experiment starts.</td>
<td>Pretest</td>
<td>A vocabulary knowledge test with 100 words where participants translate the words they know.</td>
<td>This is designed for the purpose of target word selection. Ten words that will be unknown by all participants will be integrated in vocabulary tasks</td>
</tr>
<tr>
<td>Day one</td>
<td>Vocabulary tasks</td>
<td>This is the basic experimental intervention of the study. Three vocabulary-learning tasks are designed following the model proposed by the</td>
<td>To test how learners respond to different tasks and how they benefit lexically from being engaged in performing certain tasks</td>
</tr>
</tbody>
</table>
Research questions

The present study investigates the following research questions:

1. What is the effect of task-induced involvement on incidental vocabulary acquisition and retention for learners of Arabic as a foreign language?

2. To what extent does learners’ proficiency affect how they benefit from high-involvement, vocabulary-focused tasks?

3. How does students’ performance on tasks affect their subsequent intake and retention of target words?
Method

Participants

Participants for the study were 53 English-speaking learners of Arabic (24 males and 29 females) enrolled in three intact undergraduate classes in second and third-year Arabic at an American university. In terms of proficiency, students were placed by the program through ACTFL oral proficiency exams in the range between intermediate low and intermediate high. The participants’ ages ranged from 18 to 24.

Materials

Target word selection. A major challenge in the Arabic curricula is that they were not based on frequency corpora in their vocabulary components. This made it a difficult task to predict the learners’ vocabulary knowledge based on any proficiency testing. To design a pretest, a list of one hundred words was selected from the glossary sections of the current textbooks (three levels) along with other high and low frequency items that were intuitively expected to be unfamiliar to learners. Students were required to translate the words they knew into English. This test was given to students approximately one week before the actual experiment took place. After reviewing results from the test for all participants, 10 words were selected, which were shown to be unknown to all participants. The target words and their translations are as follows: 

- aasefa ‘storm’
- aneefa ‘violent’
- yaquud ‘to drive’
- yuhawel ‘to try’
- alam ‘pain’
- taameen ‘insurance’
- tafaseel ‘details’
- dakhm ‘huge’
- taqreer ‘a report’
- mukhalafa ‘ticket’

Three vocabulary-learning tasks were designed with the ten target words embedded in each of them. Samples from the tasks are shown in Appendix A.

Experimental Tasks

Task 1. This task consisted of a text in Arabic, simplified according to participants’ average level of reading comprehension, followed by a set of ten multiple choice questions on the content of the passage. To ensure the suitability of material, the researcher collaborated with the students’ teachers, who edited the text with their students’ reading abilities in mind. The final version of the
passage (275 words) narrated a story of a car accident. The text contained ten target words bolded in the text and glossed in English in the margin. Students were instructed to read the text and answer the questions. The involvement load index for this task was posited to be 1 (Moderate Need, no Search, and no Evaluation).

Task 2. This task consisted of the same passage used in the first task but with the ten target words deleted from context. These words were given in a separate list, each followed by a translation, a brief definition and a short phrase highlighting its use in context. Students were instructed to read the text, fill in the gaps using the given words, and finally answer ten multiple-choice questions on the content of the passage, which were the same questions given to the first group. The involvement load index for this task is 2 (Moderate Need, no Search, and moderate Evaluation).

Task 3. This task involved the same list of target words, each given with a translation, a short definition and a phrase highlighting its use in context. Students were instructed to use each of the words in a sentence of their own. They were given the option of relating all the sentences to a hypothetical story of a car accident but they were not asked to combine them into a coherent text. The rationale for this option was to help less competent students build acceptable sentences. The involvement load index for this task is 3 (Moderate Need, no Search, and strong Evaluation).

Vocabulary knowledge scale (VKS)

To measure participants’ initial vocabulary learning and retention, the present study adapted Wesche and Paribakht’s (1996) vocabulary knowledge scale, which is a 5-point scale that combines written self-reported and demonstrated knowledge of new words. The scores represent the following degrees of word knowledge: 1 (never seen), 2 (familiar but meaning unknown), and 3 (meaning known but unsure), 4 (meaning provided and confirmed), and 5 (meaning is provided appropriately in a full sentence). For the immediate posttest, learners were presented with the list of ten target words and were asked to demonstrate their self-reported knowledge of each of these words on the scale provided. The same test was used for the retention test.
for the delayed posttest except that the vocabulary items were reordered. The possible range of scores was 10-50 based on students’ self-reports.

Proficiency cloze test

Because no standard written proficiency exam is available for Arabic as a foreign language, a cloze test was devised in order to establish the differences in proficiency levels between participants. To design this test, I selected a reading passage from intermediate level Arabic with a word count of 175 words. The text was developed into a cloze test in which every seventh word was deleted from context and students were required to complete the passage in any way that made it comprehensible and grammatical. Total items were 25 including both content and function words. Each answer was rated by two Arabic instructors on an acceptability scale from 1 to 4 according to the following criteria: 0 (completely irrelevant or blank answer), 1 (answer that is functionally appropriate but lexically irrelevant), 2 (answer that is lexically acceptable but functionally inappropriate), 3 (answer that is grammatically accurate and semantically equivalent to the intended meaning but could be pragmatically less appropriate), and 4 (answer that is fully appropriate grammatically and semantically).

Procedure

The experiment took place in two visits to each of the participating classes over a two-week period. The vocabulary tasks and the immediate testing were conducted on the first visit while a delayed posttest and a proficiency measure were administered during the second visit one week after the immediate testing. Practice packets were printed out and stapled so that they could be distributed randomly among students. Each packet consisted of a consent form, background questionnaire, one of the tasks, and finally the first and second posttests. On the first visit, students were asked to sign the consent forms if they were willing to participate and complete a background questionnaire. Each packet had a blank sheet with a stop
sign on it immediately after the completion of the given task. The purpose of this procedure was to have students wait for the permission of the instructor before they could proceed to the next step in the experiment. Time on task was adjusted for all the tasks to be 20 minutes maximum based on an estimation of the average duration that a low-intermediate learner would need to complete any of the three tasks.

Immediately after the time for the task was over, students were instructed to turn the page for another activity. The following page contained a distractor task where students were required to answer straightforward Arabic questions with general information about themselves. The purpose of this task was to distract students from the actual purpose of the experiment. It took about seven minutes for them to finish this exercise, after which they were asked to go to the following page where they completed the first posttest. Packets were collected that day and returned to participants on the next visit when they completed the second posttest. The delayed posttest tested the same target vocabulary except that the order of presentation was randomized. After completing the delayed posttest, participants completed the proficiency cloze test.

Analyses

The initial step in data analysis involved scoring students’ accuracy on task performance. For the reading and gap-fill tasks, the possible scores were either 1 for a correct answer or 0 for an incorrect answer. For the sentence-writing task, an acceptability rubric was developed with 1 for acceptable sentences and 0 for unacceptable sentences. Acceptable sentences should reflect understanding of the word, including structurally and semantically accurate use in a complete sentence. A half point was deducted if the sentence was comprehensible but contained more than one grammatical error. The sentence is considered unacceptable if the word is used in a minimal phrase or incomplete sentence that does not reflect understanding or the sentence contained more than three grammatical errors.
To answer the research questions, scores on immediate and delayed posttests were calculated for all groups and submitted as dependent variables to a mixed-design ANCOVA with task as a between-group, independent variable and time as a within-group factor. Proficiency scores and performance accuracy were entered as covariates to determine their significance. Effect sizes are reflected as partial eta squared ($\eta^2$), which expresses explained variance. When significant effects are found, pairwise comparisons were used to locate differences between pairs of variables. Pearson correlations were performed between proficiency, accuracy of performance, and both immediate and delayed testing scores to explore potential positive associations.

**Results**

Table 2 represents the average amounts of vocabulary learning and retention based on scores from the vocabulary knowledge scale in both immediate and delayed posttests. The descriptive statistics indicate that initial acquisition of vocabulary was the highest in the writing group, lower in the gap-fill group and lowest in the reading group. This might suggest that students who completed the writing task were more likely than other groups to recall the new target words as they encountered them in the posttest. Retention of vocabulary items followed a similar pattern with the writing group retaining the most vocabulary in the delayed posttest, followed by the gap-fill group and finally the reading group.

**Table 2:** Mean vocabulary learning scores (with SD in parentheses) for immediate and delayed posttests

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Immediate Means</th>
<th>SD</th>
<th>Delayed Means</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>17</td>
<td>17.71</td>
<td>3.26</td>
<td>14.00</td>
<td>3.14</td>
</tr>
<tr>
<td>Gap-fill</td>
<td>18</td>
<td>23.50</td>
<td>4.44</td>
<td>17.89</td>
<td>5.13</td>
</tr>
</tbody>
</table>
**Effect of task-induced involvement**

Two mixed-design ANCOVAs were conducted with task as a between-group, independent variable and time of testing as a within-group factor, and the scores of immediate and delayed posttests as dependent variables. Scores from proficiency tests, and accuracy of performance were entered as covariates. There was a main effect of task, $F(2, 47) = 21.7$, $p < .001$, $\eta^2 = .480$, and a main effect of time, $F(1, 47) = 10.21$, $p = .002$, $\eta^2 = .178$. The interaction between task and time was also significant, $F(2, 47) = 3.29$, $p = .046$, $\eta^2 = .123$. This interaction is represented in Figure 1, which illustrates that forgetting (i.e., the difference between the immediate posttest score and the delayed posttest score) is larger in the sentence-writing and gap-fill groups than in the reading group. A Bonferroni post hoc test showed significant mean differences between the reading task and the gap-fill task ($p = .001$) and between reading and sentence writing ($p < .001$), while the mean difference between the sentence writing and gap-fill tasks approached significance ($p = .053$).

**Figure 1**: Means plot for the interaction between task and time effects
Effect of proficiency

The second research question is concerned with how proficiency score, as a covariate, affected learners’ performance and subsequent vocabulary acquisition from given tasks. To determine proficiency scores for all participants, the objective cloze test results were compared. A One-Way ANOVA showed no significant differences among groups in terms of proficiency scores, F (2, 52) = 2.97, p=.071. Across participants, the mean proficiency score was 25.8 out of a maximum score of 50. Mean proficiency scores for all groups are shown in Table 3.

Table 3: Mean L2 Arabic proficiency scores (with SD in parentheses) for the three groups

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Means</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>17</td>
<td>23.24</td>
<td>5.83</td>
</tr>
<tr>
<td>Gap-fill</td>
<td>18</td>
<td>25.06</td>
<td>7.70</td>
</tr>
<tr>
<td>Sentence writing</td>
<td>18</td>
<td>29.00</td>
<td>8.37</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>25.81</td>
<td>7.67</td>
</tr>
</tbody>
</table>

Results of the mixed design ANOVA revealed a significant effect of proficiency on vocabulary learning and retention, F (1, 47) = 13.29, p=.001, η2=.213. This means that, in addition to the effect of task type, proficiency levels also contributed to how much learners benefited from given tasks in terms of vocabulary learning.

Effects of task performance accuracy

The third research question asked whether learners’ performance on tasks affected the way they benefited from them. To determine performance accuracy across participants, two raters scored learners’ responses to task items on a 10-point scale. The
mean accuracy scores are shown in Table 4. A One-Way ANOVA showed no significant differences among the three groups in accuracy scores, F (2, 52) = 1.61, p=.211. This means that, to the extent that accuracy is reflective of task difficulty, the three tasks were roughly as difficult.

Table 4: Mean accuracy scores (with SD in parentheses) for the three groups

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Means</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>17</td>
<td>7.18</td>
<td>1.47</td>
</tr>
<tr>
<td>Gap-fill</td>
<td>18</td>
<td>6.72</td>
<td>2.49</td>
</tr>
<tr>
<td>Sentence writing</td>
<td>18</td>
<td>7.94</td>
<td>2.07</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>7.28</td>
<td>2.09</td>
</tr>
</tbody>
</table>

Given that the three tasks were different in their demands, we were more concerned with how task performance was effective as a covariate with task type. The mixed design ANOVA showed that the effect of accuracy of performance was significant, F (1, 47) = 5.79, p=.020, η²=.110. This means that the amount of words acquired and retained was influenced to some extent by how well participants completed the task in the first place. A follow-up investigation revealed a significant correlation between proficiency and task performance across tasks (r= 0.314, p=.022).

Looking at task performance within groups, it showed that vocabulary learning did not correlate with performance in the reading task (r= .092, p=.724), or the gap-fill task (r=.072, p=.777), but that there was a significant correlation with sentence writing accuracy in the immediate posttest (r= .710, p < .001) and the delayed posttest (r=.502, p =.034). Across all participants in all three tasks, a
significant correlation was only found between accuracy scores and immediate posttest scores ($r = .310$, $p = .024$). The scatter plots in Figures 3 and 4 display the correlations between task performance and both immediate and delayed posttest scores with different symbols for the three groups.

*Figure 2: Correlations between task performance and immediate post test scores*
Figure 3: Correlations between task performance and delayed post test scores

Visual comparison of the scatter plots suggests that the patterns in the immediate and delayed tests are closely similar. We can also see a gradually rising line for the writing group, while the distribution for the other two groups is not consistent. Across all participants, the scatter plots do not show a positive correlation in either the immediate or the delayed posttests.

Discussion

Main results of the study confirmed the predictions of the hypothesis in that high-involvement tasks produced more vocabulary learning than less-involving tasks, as determined by the involvement index. In particular, the sentence-writing task produced higher scores of acquisition and retention of vocabulary, followed by the gap-fill task, while the least learning came from the reading task. These results, then, support the claims made by Hulstijn and Laufer (2001)
that incidental acquisition of vocabulary is supported by the degree of involvement with the target words induced through specifically designed tasks. However, the interaction between task time effects indicated that the groups of gap-fill and sentence writing showed a decrease in retention over time, which means they forgot what they initially picked up from task performance.

The second research question examined the effect of proficiency levels on the students’ ability to benefit from task performance. One contribution of this study is introducing a novel measure for Arabic language proficiency. Using a specifically designed cloze test, I was able to estimate ratings for students’ linguistic abilities. Although no significant proficiency differences were found between groups, minor differences across participants did exist, and these differences had a significant effect on how much learners benefited from the learning tasks. In general, the present study provided empirical evidence that proficiency differences, though minor, had a significant effect on how learners benefitted from tasks with different involvement loads. This could be explained in terms of learners’ use of their linguistic resources in completing different tasks. It is reasonable to assume that learners with different language abilities would approach linguistic tasks differently. This can have direct implications for task difficulty and complexity, and poses a critical question as to the criteria that are set in empirical studies when designing tasks and comparing their relative effectiveness on learners’ performance and development.

The third research question explored the relationship between learners’ performance accuracy on given tasks and their vocabulary learning outcome. In line with the finding that learners’ proficiency reflected on their task performance and reflected on their vocabulary learning, this study also found that accuracy of performance correlated with learners’ scores in the immediate posttest (but not with the delayed test). Accuracy scores were a significant covariate with regard to learners’ vocabulary acquisition. This means that the more accurately students performed their tasks; the more they were engaged with them and the more words they picked up incidentally. This could be explained by the assumption
that, besides the task-induced involvement, a portion of involvement may be driven by the learners’ self-engagement with the task and their intent to do it well. The effect of this learner-induced involvement was temporary, as shown by the absence of a significant correlation between performance accuracy and vocabulary learning after one week.

An examination of the scores of each group separately revealed that only performance within the writing group correlated significantly with their vocabulary learning results. One explanation can be drawn from the cognition hypothesis (Robinson, 2003; Robinson and Gilabert, 2007) which maintained that increasing the demands of L2 tasks could lead to differential language use, depth of processing for input, and more attention to output. This subsequently encourages further attention to the input given in task-based activities and helps in more retention of target items for subsequent use. In general, terms; the cognition hypothesis supports the view that task complexity affects task performance, which eventually feeds into acquisition and retention. These assumptions simply support the results of the study on the effects of performance accuracy and the superior effects of certain tasks, like sentence writing.

To sum up, the findings of the present study validated the Involvement Load Hypothesis in its current formulation by replicating Laufer and Hulstijn’s (2001) basic empirical design. However, the present study did not find evidence to support the claim that the involvement load is an overriding factor in task-based vocabulary learning. Learners’ proficiency levels and accuracy of their task performance were shown to be contributing factors that interact with task type in enhancing incidental vocabulary acquisition and retention.

**Directions for future research**

The involvement load hypothesis evolved as a means of explaining the construct of incidental vocabulary acquisition. It based its claims on the design of specific tasks that arguably encourage incidental learning. In general, the features of tasks targeting
incidental learning are usually debatable in the literature, leaving room for interpretation of the learning outcomes of vocabulary-focused tasks. Further research efforts are needed to implement standard guidelines in designing and conducting learning tasks. A related controversy in the light of the hypothesis is the type of tasks being implemented for validating its claims. The comparison between reading comprehension and gap fill on the one hand, and sentence writing on the other hand, is actually a comparison between input-oriented tasks and output-oriented tasks. The advantage of output could by itself bias the results towards sentence writing, not only due to the cognitive load but also due the effect of language production on acquisition (Swain, 1995). This calls for another consideration of the distribution of the components underlying involvement (need, search, evaluation) for more robust evidence to validate the Involvement Load Hypothesis.

Another argument that could be made in this regard is that the construct of involvement was empirically tested in one distribution pattern of components, with need and search values fixed and only evaluation varied (no evaluation, moderate evaluation, and strong evaluation). The pattern of results, then, in previous studies and the present study alike could be due to this variation. One other issue with the hypothesis concerns the component of search, which is operationalized as either absent or present. This could be problematic, since it does not match with the degrees of the other components of need and evaluation, which follow a three-point scale. One suggestion to refine this component is to distinguish three types of search: looking up words in a dictionary, searching for word meanings in a given glossary or inquiring about meaning with a teacher or classmate. Empirical investigation would be recommended to determine the variable involvement loads of different modes of search.

One final argument that could be relevant to this research area is related to the distinction between incidental and intentional learning in vocabulary studies. Some previous studies have almost adopted Hulstijn’s (2001) distinction that incidental learning applies when learners are not forewarned of a vocabulary test after the
experiment. However, there seemed to be a lot of terminological vagueness when referring to incidental learning in most recent studies. In the involvement load hypothesis studies, the reference to the type of learning as ‘incidental’ was cautiously presented and at most times ignored, even though the hypothesis evolved as a means of interpreting the results of incidental vocabulary acquisition studies. The type of tasks used in the present study may invite some arguments as to whether they only triggered incidental acquisition, driven by the fact that we cannot know for sure what cognitive processes occurred in the mind of learners while they performed the tasks. There may be certain elements in the design of a certain task that draw more than an incidental attention to the target words. Paper-and-pencil studies cannot provide full insight into what learners actually do when they engage with novel lexical items within task performances. One promising area of research is to employ online measures to allow for tracking the cognitive processes involved in incidental learning and improve our understanding of what actually constitutes incidental acquisition.

Conclusion

The importance of this study lies in its exploration of tasks and materials in Arabic language instruction as a recently evolving field of study. One unique contribution of the study is its attempt to devise new materials and adopt new measures to meet language-specific needs and develop research instruments in the field. The study validated the Involvement Load Hypothesis in its current form and empirical design. However, the investigation of other factors also pointed to some ways to refine the components underlying involvement and to modify some of the theoretical underpinnings of the construct.

This study enables teachers and practitioners in the field to make informed choices about how to create meaningful opportunities for “incidental” vocabulary learning in the classroom. These can complement the usual practice of teaching vocabulary explicitly, which has been shown to be more reliable and durable (Laufer, 2005; Schmitt, 2008). One other point to take away from this
study is the need to establish what constitutes a task in a classroom setting. Researchers in this area should relate to the actual practice of teaching and integrate relevant designs. For example, they may need to question whether a specific kind of reading task or gap-fill task is commonly used in particular second or foreign language class activities or teaching materials. Based on that, they can empirically test the efficiency of these practices and bring to bear their own SLA perspectives on certain teaching practices. In this way, teachers are more likely to be informed by research and can more easily apply strategies that induce engagement with vocabulary while considering other factors of exposure, attention, rehearsal, and individual differences.

As the results of the present study pertain to English speaking learners of Arabic, a lot of population-specific insights can be drawn from the study procedures. The average vocabulary outcome for all participants is relatively low and may reflect difficulties on the part of English-speaking learners in acquiring new words. One plausible reason behind that can be related to the unstable and somewhat variable vocabulary sizes of learners at this stage. The textbook, as the major source of their vocabulary knowledge, is believed to be a contributor in their unbalanced and insufficient lexicon. Additionally, learners did not get focused training in the root-and-pattern system, which constitutes a critical and consistent component in Arabic lexicon and morphology. Teachers and practitioners believe that, although Arabic lexicon has been shown to be rich and complex, a sound training in the root-and-pattern system can offer a remarkable boost in vocabulary building for learners, given that the Arabic root system is highly predictable and demonstrates a mathematical regularity. No considerable research has been done in Arabic vocabulary learning, particularly the factors that enhance or slow the incremental nature of lexical acquisition.

A research-pedagogy collaborative effort is essential to advance the field of Arabic teaching. Curriculum practitioners should be informed by corpus research in developing task-based modules that target high frequent vocabulary and recycle them through theme-
based topics for enhancing communicative fluency. Special attention should also be given to testing instruments and standard proficiency measures that can serve research purposes, curriculum design, and program evaluations. Both quantitative and qualitative approaches are important for investigating the current state of Arabic language instruction and acquisition as evolving fields in American universities. This trend is likely to provide implications for developing practices in Arabic language pedagogy and encourage further research in the field.
References


Appendix A: Vocabulary learning tasks

Task 1. Read this text and then answer the questions on the following page: you are not allowed to look at the questions before you read the whole text.

1. to drive
2. storm
3. huge
4. to try
5. violent
6. fell down
7. details
8. report
9. pain
10. insurance

1. to drive
2. storm
3. huge
4. to try
5. violent
6. fell down
7. details
8. report
9. pain
10. insurance

<table>
<thead>
<tr>
<th>1. to drive</th>
<th>حادثة سيارة</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. storm</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>3. huge</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>4. to try</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>5. violent</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>6. fell down</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>7. details</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>8. report</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>9. pain</td>
<td>حادثة سيارة</td>
</tr>
<tr>
<td>10. insurance</td>
<td>حادثة سيارة</td>
</tr>
</tbody>
</table>

استيقظ خالد من النوم متأخرا. غضب من زوجته لأنها لم توقظه مبكرا، ولكنها قالت له أنها كانت متعبة طوال اليوم وظلت أنه أيضا متعب من السفر ليلة أمس ولم توقظه. أعادت له الإفطار فتناوله بسرعة ولم يستطع أن يأخذ الأولاد معه إلى مدرستهم لأنه كان عنده معود وخشى أن يتأخر عنه. وأراد أن يذهب إلى عمله. استعد وليست ملامسه وخرج بسرعة وركب السيارة. كانت السيارة تحتاج إلى إصلاح ولكنه تجاهل ذلك وقال أنه ليس عنده وقت. وبها القيادة وكان صوت السيارة غريبًا ولكنه استمر في القيادة ودخل إلى الطريق السريع. وكان يقود سيارته في الطريق السريع عندما بدأت عاصفة شديدة وأمطار كثيرة. لم يهتم بذلك لأنه معتاد على القيادة في ظروف صعبة. فجأة شاهد سيارة ضخمة أمامه ولم يستطع أن يتوقف. حاول أن يذهب بعيدا ولكنه فشل فاصطدم بها. وكانت حادثة عنيفة. انكسر زجاج سيارة السائق الآخر وسقط خالد من السيارة على الأرض. أتصل السائق الآخر بالبوليس ليبلغ عن الحادث. جاء رجل البوليس بسرعة وسلمهما عن تفاصيل الحادث ليعرف من الذي أخطأ وما هو سبب الحادثة. روى خالد للشرطي كيف كانت الحادثة. كتب الشرطي تقريرًا عن الحادث وأعطى خالد مخالفته لأنه كان يقود بسرعة وأخذه إلى المستشفى لأنه كان يعاني من ألم شديد في رجليه. اتصل السائق الآخر بشركة التأمين ليعرف كيف يمكن أن يصلح سيارته. عندما علمت زوجة خالد بالخبر ذهبها فورا إلى المستشفى لتلتزم على زوجها وتركت أولادها مع جدتهم. قال الطبيب أن خالد يحتاج أن يمكث في المستشفى ثلاثة أيام على الأقل حتى تتأكد أن كل شيء على ما يرام. واتصل خالد بصاحب العمل ليطلب منه إجازة لمدة أسبوع حتى يستطيع أن يرجع إلى العمل.
الأسئلة

Choose the answer that completes each sentence

1- لماذا أعطى رجل البوليس مخالفة لخالد؟
أ) لأنه كان يقود بسرعة
ب) لأنه كان هو سبب الحادثة
ج) لأنه كان يعاني من ألم في رجله

2- لم يأخذ خالد الأولاد معه لأنه 
أ) كان يوم أجازة
ب) كانوا يتناولون الإفطار
ج) كان متاخرًا عن موعده

3- كانت الحادثة عنيفة بسبب 
أ) البوليس جاء متأخرا
ب) العاصفة والأمطار
ج) السيارة تحتاج إلى إصلاح

4- لماذا سقط خالد من السيارة?
أ) خالد يحتاج إلى عملية
ب) خالد يبقى في المستشفى قليلا

5- لماذا أعطى رجل البوليس مخالفة لخالد؟
أ) لأنه كان يقود بسرعة
ب) لأنه كان هو سبب الحادثة
ج) لأنه كان يعاني من ألم في رجله

6- لم يهتم خالد بإصلاح السيارة لأنه
أ) لا يعرف كيف يصلاحها
ب) معتاد على القيادة في ظروف صعبة
ج) ليس عنه وقت

7- لماذا أعطى رجل البوليس مخالفة لخالد؟
أ) لأنه كان يقود بسرعة
ب) لأنه كان هو سبب الحادثة
ج) لأنه كان يعاني من ألم في رجله

8- لم يهتم خالد بإصلاح السيارة لأنه
أ) لا يعرف كيف يصلاحها
ب) معتاد على القيادة في ظروف صعبة
ج) ليس عنه وقت

9- كم يوما طلب خالد من صاحب العم إجازة؟
أ) لمدة ثلاثة أيام
ب) لمدة أسبوع
ج) لمدة شهر

10- الطبيب في المستشفى قال إن
أ) خالد يحتاج إلى عملية
ب) خالد يبقى في المستشفى قليلا
ج) خالد لا يقود السيارة مرة أخرى

5- أراد رجل البوليس أن يسمع

أ) تفاصيل الحادث
ب) تقرير المستشفى
ج) صوت السيارة

6- اتصل السائق بشركة التأمين لأنه

أ) يعاني من ألم في رجليه
ب) يريد أن يصلح سيارته
ج) يريد أن يأخذ إجازة
Task 2. Complete the story of a Car accident using the given word list in the next page then answer the following questions.

### Car accident (حادثة سيارة)

Asteqofa al-dakhil min al-nom muta'akmara. Ghasab min zujotuhu lana lami litoqofe mibka wa lakena qaallta leh.

Anhaa kamaa matamuba thawal al-yum wa ootinti anhaa ammaa mutub min al-safir liyabta ams fum litoqof. Aqadit
leh al-effatat futawlah buseruwa wa lam yistemu an yaxh al-a'dala muhe leh muddaristi lana lami kawn unhe.

Mowood wuxashi an yitaqar yu ne. Wa arrad aan jibebi leh unh. Asaadd wa lami leh wargi
buseruwa warqib al-siyara. Kana al-siyara zharji la la ila al-islaahl baleh ne tajebi leh wata al-lando ne yus
unhe qof. Wabda laqiadi wa kana cusut al-siyara griiba baleh ne astu'm laqiadi wad chum leh saal.

Al-siyara bi al-sirr. Wa kana .....................

Sheidiwa amebata kohore. Lam yehem biilik la ila la mdu laqiadi la ila qof
\(\text{صعبة}. \) Fqaha shahid al-siyara. ............. Amamne la fustuq an yitqof.


An nakshar zajad al-siyara al-sa'iq al-a'xwar wa .....................

Halad min


al-balwis buseruwa wa sa'lahaqna ila. ..............

Handith layefur min an bi'ti aqada wa ma hoo


En al-handith a 'aqi al-ma'afiika lanee kaen yeydem buseruwa wa xadhe al-mu'stefi lakaan

Alriganiy m ..................... Shebdi fi rejhi. Etsal sa'iq al-a'xwar bisharika

Layefur kif yemka an yisalab siyara. Amda a'malta zujoda halad bal'hihi dahamta fara ila

al-mu'stefi limdunan ila zujoda wa tarkit a'qada baleh nhad dib. Qal al-matibb an halad yihatඹ

An yekht in al-mu'stefi limdan sheb yi'amila ila 'aqada baleh nhad dib akki sah la yamama. Wata'sal

Halad bi sahhib al-umal liyebbi mina ilajada limdan ashub biqi iliy yemtstuq an yirjub ila al-umal.
<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>تفاصيل</td>
<td>Details of something; we say تفاصيل الخبر – تفاصيل القصة</td>
</tr>
<tr>
<td>سقط</td>
<td>Fell down; we say سقط على الأرض – سقط من الدور الثاني</td>
</tr>
<tr>
<td>ألم</td>
<td>Pain or suffering; we say أشعر بالألم في رأسى</td>
</tr>
<tr>
<td>يقود</td>
<td>To drive a car, truck or any means of transportation; we say لا يجب أن نقود السيارة بسرعة</td>
</tr>
<tr>
<td>عاصفة</td>
<td>Storm or a strong gust of wind; we say هناك عاصفة شديدة متوقعة غدا</td>
</tr>
<tr>
<td>حاول</td>
<td>To try or to attempt at; we say حاول أن ينام ولكنه لم يستطيع</td>
</tr>
<tr>
<td>التأمين</td>
<td>Insurance or securing; we say نحتاج إلى تأمين المكان لزيارة الرئيس</td>
</tr>
<tr>
<td>ضخم</td>
<td>Huge or enormous; we say هذا بيت ضخم</td>
</tr>
<tr>
<td>عنيف</td>
<td>Violent/severe; we say ألم عنيف</td>
</tr>
<tr>
<td>تقرير</td>
<td>A report; we say تقرير عن الحادث</td>
</tr>
</tbody>
</table>
الأسئلة

Choose the answer that completes each sentence

1- كان خالد يقود سيارته بسرعة لأنه

أ) استيقظ متأخرا وكان عنده موعد مهم
ب) كان غاضبا من زوجته
ج) كان يريد أن يتناول الإفطار

2- لم يأخذ خالد الأولاد معه لأنه

أ) كان يوم أجازة
ب) كانوا يتناولون الإفطار
ج) كان متاخرا عن موعده

3- كانت الحادثة عنيفة بسبب

أ) البوليس جاء متأخرا
ب) العاصفة والأمطار
ج) السيارة تحتاج إلى إصلاح

4- لماذا سقط خالد من السيارة؟

أ) لأنه حاول أن يذهب بعيدا
ب) لأن زجاج السيارة انكسر
ج) لأن السيارة الضخمة صدمته

5- أراد رجل البوليس أن يسمع

أ) تفاصيل الحادث
ب) تقرير المستشفى
ج) صوت السيارة

6- اتصل السائق بشركة التأمين لأنه

أ) يعاني من ألم في رجله
Task 3. Use the given words in sentences of your own. You have the option to make your sentences describe a situation of a car.

<table>
<thead>
<tr>
<th>Word</th>
<th>Sentence in Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>تفاصيل (n.pl)</td>
<td>تفاصيل الخبر – تفاصيل القصة</td>
</tr>
<tr>
<td>سقط (v.)</td>
<td>سقط على الأرض – سقط من الدور الثاني</td>
</tr>
<tr>
<td>ألم (adj.)</td>
<td>أشعر بالألم في رأسي</td>
</tr>
<tr>
<td>يقود (v.)</td>
<td>لا يجب أن نقود السيارة بسرعة</td>
</tr>
<tr>
<td>عاصفة (n.)</td>
<td>هناك عاصفة شديدة متوقعة غدا</td>
</tr>
<tr>
<td>حاول (v.)</td>
<td>حاول أن ينام ولكنه لم يستطيع</td>
</tr>
<tr>
<td>التأمين (n.)</td>
<td>نحتاج إلى تأمين المكان لزيارة الرئيس</td>
</tr>
<tr>
<td>ضخم (adj.)</td>
<td>هذا بيت ضخم</td>
</tr>
<tr>
<td>عنيف (adj.)</td>
<td>ألم عنيف</td>
</tr>
<tr>
<td>تقرير (n.)</td>
<td>تقرير عن الحادث</td>
</tr>
</tbody>
</table>