The Interaction between Involvement Load Hypothesis Evaluation Criterion and Language Proficiency: A Case in Vocabulary Retention

Hassan Soleimani

Payame Noor University, Tehran

h_soleimanis@pnu.ac.ir

Ali Asghar Rostami Abu Saeedi

Payame Noor University, Tehran

rostamiabu110@yahoo.com

Abstract

A long-standing debate among the vocabulary researchers is the depth of processing to learn vocabulary. This paper is a quantitative research which considers a revision in the "involvement load hypothesis" proposed by Laufer and Hulstijn in 2001. It investigates the role of proficiency and *evaluation* in this hypothesis in order to better reveal its potential contribution to vocabulary learning. It was based on *task-induced involvement* that compares different tasks in incidental vocabulary acquisition in EFL context. The participants were 66 learners from two different English institutes who were classified into two major high and low proficient groups based on Nelson Proficiency Test. The participants in each group were randomly assigned to three tasks prepared to compare "moderate", "strong", and "no evaluation" in involvement load hypothesis. The "strong evaluation" subgroup (making original sentences) in low proficiency supported Laufer and Hulstijn's hypothesis and yielded better retention of the target words. The study suggests that the level of proficiency and evaluation in task induced involvement needs reconsideration. The results have implications for language teachers, materials developers, and syllabus designers.

Keywords: Involvement Load Hypothesis, Task-Induced Involvement, Evaluation, Proficiency, Involvement Index

1. Introduction

Since the rise of cognitive approaches to learning which gave prominence to memory and knowledge in language learning, some information processing models of memory have emerged. In 1972, Craik and Lockhart suggested their *depth of processing hypothesis* as an alternative to the three-store model. It holds that elaborate processing of unfamiliar words is the key factor in better retention of them. The claim which often relates to models of memory is considering attention as the necessary and sufficient perquisite for long-term retention (Schmidt, 1994). His suggestion of *noticing hypothesis* along with *comprehensible output* by Swain (1985) made Laufer and Hulstijn (2001) provide a solution to the instruction of vocabulary, a long-standing debate among the vocabulary researchers. Their innovative construct which is a branch of incidental learning is called "involvement load hypothesis". It states that by changing the amount of its constituent parts, cognitive *(need, search)* and motivational *(evaluation)* factors, the retention of unfamiliar vocabulary items will be affected. Although the involvement load hypothesis was first suggested in 2001, there are not enough studies (e.g., Cheng, 2011; Keating, 2008; Kim, 2008; Xu, 2009; Walsh, 2009) to evaluate the theory. Kim (2008) advocates the feasibility of this theory but is suspicious of its components equal importance. In addition, the results of some studies to investigate the effect of the level of proficiency (Keating, 2008; Kim, 2008; Xu, 2009) in different types of tasks in involvement load hypothesis on vocabulary acquisition are somehow contradictory. The present study is determined to reveal these points in involvement load hypothesis using tasks with different involvement indices.

2. Review of Related Literature

Cognitive psychology holds the information processing as its heart. This processing model of memory relates to attention, encoding, elaboration, storing, and retrieving. As being foolproof is not the attribute of any human theory, Craik and Lockhart (1972) mention some points to refer to the partiality of this model and in effect present the basis of their depth of processing hypothesis on processing stages. Their hypothesis states that the factor which affects long-term retention is not the length of keeping information in primary memory, but the depth of processing it. They have also suggested several levels of processing depth, structural features of lexical items are processed at a shallower level than semantic features. As with any theory, Craik and Lockhart's (1972) hypothesis has some strengths which Hulstijn and Laufer (2001) maneuver on. Their involvement load hypothesis states that retention of target words depends on nothing but the involvement index of any task (Laufer & Hulstijn, 2001). Although depth of processing hypothesis did not get through more than three levels, the involvement load hypothesis in accordance with Solso (1988) "starting with shallow sensory analysis, and proceeding to deeper, more complex, abstract, semantic analysis" (Solso, 1988, p. 133). Laufer and Hulstijn (2001) proposed task induced involvement based on it. In fact, in their attempt to make vocabulary learning more efficient, they have taken in to account artificial tasks like gap-filling or summary writing that can be designed in such a way to manipulate their involvement load. They emphasize that "This proposal should be conceived as a first attempt to stimulate researchers as well as practitioners to operationalize the general labels of 'attention' and 'elaboration' in to concrete task-specific constructs" (Laufer & Hulstijn, 2001, pp. 13-14). Their involvement load hypothesis claims that by changing the amounts of its constituent parts, cognitive (need, search)

and motivational (evaluation) factors, the retention of unfamiliar vocabulary items will be affected. Each of these components has a "moderate" and a "strong" value. The evaluation factor can be moderate if it makes learners assess appropriate word for a specified context and strong if in addition it makes them create an original sentence. On the other hand, the need component can be either extrinsic like when it is needed to use in a sentence because of teacher's request or intrinsic like when it is self-imposed by the learner. Search is concerned with noticing to the form-meaning relationship of the target words (Schmidt, 1994, 2000). Trying to find the form or meaning of unfamiliar L2 vocabulary items using a major source of information such as a dictionary or teacher is what Laufer and Hulstijn (2001) meant by using this component.

Laufer and Hulstijn's (2001) study has been considered as the first attempt that linked the abstract concepts of noticing, elaboration, motivation, or need at the low level of task, which resulted itself in subsequent studies. Kim's (2008) study can be regarded as one of them which investigated two different levels of proficiency in 20 adult English as L2 learners. In fact, he found no significant difference between the two levels of proficiencies in doing the tasks. Besides, Kim (2008) did not yield the same value to the motivational and cognitive components of the involvement load in vocabulary learning and knew the evaluation component of a higher superiority than search and need. In another study, Xu (2009), who did his experiment with 152 freshman university students, found that level of proficiency of the students is important but only for the immediate task types.

3. Research Questions

The purpose of the present study was to examine whether using different tasks can enhance vocabulary acquisition of high and low proficient Iranian EFL students. Our specific research hypotheses were:

H01: The task-induced involvement has no effect on the retention of low proficient Iranian EFL students

H02: The task-induced involvement has no effect on the retention of high proficient Iranian EFL students.

4. Method

To meet the requirements of the study and find answers to the questions formulated in the present study, following Mackey and Gass (2005), a quasi-experimental research with comparison groups design was conducted with non-random sampling technique.

5.1. Participants

The participants in the present study were 66 Iranian EFL male and female students between 19-25 years old who studied at intermediate level in two branches of English language institutions. Two language institutions were used since each institution could only provide three intact classes containing 11 students that were not enough for the experiment. The participants had received English training for about three institutional terms and had the same core and supplementary materials. The research was conducted during their normal class time during approximately a six-week period of a semester.

5.2. Instruments

In order to conduct our research, we resorted to some materials. In the following, we will explain the instruments which we applied for data collection: Nelson Placement Test, a reading text along with five reading comprehension questions, and target vocabulary items.

5.2.1. Nelson Placement Test

Since we had our sample population from two institutions, we were not safe enough to use the students' previous term scores. Therefore, for a more clear-cut assignment of the participants into two major groups, the second version of Nelson Placement Test (intermediate 200 A) which contained 50 multiple-choice items was administered.

5.2.2. Reading Passage

In order to choose an appropriate reading comprehension passage, three intermediate texts were chosen based on their similar readability index, being within the students' general knowledge and vocabulary domain, adapting five multiple choice reading comprehension questions from it in such a way that answering them required understanding the target words in order to let us have the "moderate need" component of Laufer and Hulstijn's (2001) innovative construct. In addition, each target word had to occur only once to attribute the results to Laufer and Hulstijn's and not to multiple exposures known by some researchers (e.g., Nation, 1990) as one of the factors playing a role in learning from context.

After selecting the most comprehensible text by the participants' teachers and intermediate students of the same institutes who were not the participants

of our study, the text which had already been used by Walsh (2009) in a research investigating the same variables was chosen. It was about "child labor" and the ways of decreasing this problem. The reading text contained 326 words with a Gunning Fog index readability of 7.73 using an on-line utility operating system. Furthermore, the reading text in this study had some L1 or L2 glosses in the margin which in accordance with some researchers (i.e., Craik & Lockhart, 1972; Craik & Tulving, 1975; Laufer & Hulstijn, 2001; Nation, 2001) contribute to a stronger relationship between form and meaning and thereby enhance the chance of remembering them.

5.2.3. Target Vocabulary Items

Intermediate students of the considered institutions, the participants' teachers, and the person who was responsible for placing the students in the proper level were asked to select the unknown words of Walsh reading passage for our sample population. All of the vocabulary items which they selected composed the pretest items which was administered to determine the words which our intermediate participants did not know. They were required to write the L1 or L2 equivalents of the words in the pretest which had 36 items. Afterwards, its reliability coefficient was calculated as .62. Finally, ten words in the text including two verbs, six nouns, one adjective, and one adverb were selected: *plantation, fair, demonstrations, crops, sweatshop, and fiber, partly, blame, march,* and *shrimp.*

5.3. Data Collection Procedure

Thanks to Nelson Placement Test we could have an accurate classification of our sample population into two major high and low proficient groups. In line

with the purpose of the experiment, we randomly categorized each major group into three groups in order to assign each of them to one of the three task conditions, which had their only difference in evaluation component of task-induced, in their normal class time period. In order to be confident about the students' lack of any prior knowledge of the target vocabularies a pretest was administered. One week later, the students were given the treatment of tasks. In parallel with incidental learning lines, answering the five multiple-choice questions was explained as the purpose of the experiment and nothing was told them of the posttests. To put it simply, we tried to decrease the students' affective filter and diminish their responsibilities when they encountered an unknown word. Immediately after the completion of the tasks, the subjects' papers were collected to avoid if any subject would memorize the target words for further searching later. Subsequently, learners were required not to talk or discuss the target words after the papers were collected. This was done because peer exchanges can manipulate need, search, and evaluation components of the hypothesis which was not in line with our research purpose (Walsh, 2009).

5.3.1. Tasks

Blanked out reading text with multiple-choice questions were prepared for the participants in Task A to fill in the blank spaces using the ten L2 marginal glosses and answer the questions. The numerical involvement load of this task was two since it induced a moderate need, no search, and "moderate evaluation" on the participants ($1+0+1=2$). In Task B the students were required to read the same reading text, answer the same questions, and write original sentences with L1 glossed target vocabularies. The involvement index is calculated as three since a moderate need, no search, and "strong evaluation" are induced on the participants ($1+0+2=3$). Having the placebo role, the

participants of Task C had to do nothing except answering the multiple-choice questions. The involvement index for this task was one since it had a moderate need, no search, and no evaluation on the part of students (1+0+0=1).

5.4. Data Analysis

The scoring procedure of the vocabulary immediate and delayed posttests was one point for correct equivalents, half a point for semantically close equivalents, and zero point for wrong equivalents. The homogeneity of the six groups was investigated on the participants' scores on Nelson Placement Test. Based on meeting the normality assumption, parametric and nonparametric data analysis procedure was decided on. The data were yielded to SPSS (version 18.0) for analysis. A significance level of .05 was assumed for the results.

6. Results
6.1. Testing Normality

Three numerical and graphical normality indices of skewness, One-Sample Kolmogorov-Smirnov test, and Q-Q plot were run in order to reveal if the data were normally distributed. Since the normality presumption was met in the first hypothesis, one way $ANOVA$ was conducted. Regarding the homogeneity of variances, the Levene Test of homogeneity of variances made it clear that the variances of the three groups were significantly different at the $p < .05$ level. Therefore, the assumption of homogeneity of variances was violated: $F(2, 96) = 10.92$, $p = .000$ and nonparametric statistical test Kruskal-Wallis was run.

On the contrary, in the second hypothesis the normality assumption was not met; therefore, the nonparametric statistics Kruskal-Wallis was applied in order to test the hypothesis.

6.2. Homogeneity Test Results

The homogeneity of all groups in the study was examined based on their scores in Nelson Placement Test. The K-S normality index made it clear that the six groups met the assumption of normality: $p=.200$ at the p<.05 level. Therefore, the parametric statistic test one-way *ANOVA* was conducted. This test was done to confirm us that there was no statistically significant difference between the means of the six groups before administering the tasks. As Table 1 demonstrates, Levene Test of homogeneity of variances clarified that the variances of scores across the groups were homogeneous, $F(5, 60)=.19$, $p=.96$ (2-tailed) at the p<.05. Therefore, it can be found that the six groups were homogeneous and belonged to the same population.

Table1. *Test of Homogeneity on the Proficiency Scores*

Levene Statistic	df1	df2	Sig.
.19	5	60	.96

6.3. Findings of the Study

This study was attempted to investigate the following two hypotheses.
H01: The task-induced involvement has no effect on the retention of low proficient Iranian EFL students.

This research hypothesis was about comparing different tasks with different involvement indices on the retention of vocabulary items in low proficient learners. The non-parametric test of Kruskal-Wallis was conducted for comparison of the means, which can be seen in Table 2 and Table 3, that

showed a statistically significant difference at the $p<.05$ level in the vocabulary retention of the three groups, χ^2 (2, $n=99$)$=17.11$, $p=.000$. Therefore, it is implied that the first null hypothesis was rejected.

In other words, there was a statistically significant difference in vocabulary mean scores of the low proficient group who did the three tasks. Table 4 demonstrates that the sentence making group recorded a higher median (*Md*= 5) than the other two groups of gap-filling and control with values of 4 and 2 respectively.

Table 2. *Descriptive Statistics of the Vocabulary Tests of Low Proficiency in Terms of Ranks*

	Tasks	N	Mean Rank
dimension1	A	33	58.42
	B	33	58.18
	C	33	33.39
	Total	99	

Table 3. *Kruskal Wallis Test of the Vocabulary Tests of Low Proficiency*

	tests
Chi-square	17.11
Df	2
Asymp. Sig.	.000

Table 4. *Median of the Three Groups on the Vocabulary Test of Low Proficiency*

	Tasks	N	Mean Rank
Dimension 1	A	33	4.00
	B	33	5.00
	C	33	2.00
	Total	99	3.00

In the post-hoc test of Mann-Whitney U, which required a Bonferonni adjustment to control for Type errors, the alpha level of.05 was divided by 3 (the number of pairs to be compared). Regarding this stricter level of p<.017, Tables 5 and 6 demonstrate that the results of the first Mann-Whitney U Test forthe retention mean score of gap-filling with L2 glossing group (*Md*=4, *n*=33) and writing original sentences with L1 glossing (*Md*=5, *n*=33), U=541.50,z =-.039, p =.969, d=.004 were not significantly different with a very small effect size in accordance with Cohen (1988) criteria of 0.0-0.2=small effect, 0.3-0.5=medium effect, and 0.6-0.9=large effect.

Table 5. *Mean Ranks of the Gap-Filling with L2 Gloss and Sentence Making with L1 Gloss Groups of Low Proficiency*

	tasks	N	MeanRank	Sum of Ranks
dimension1	A	33	33.41	1102.50
	B	33	33.59	1108.50
	Total	66		

Table 6. *Mean Ranks of the Gap-Filling with L2 Gloss and Sentence Making with L1 Gloss Groups of Low Proficiency*

	tests
Mann-Whitney U	541.50
Wilcoxon W	1102.50
Z	-.039
Asymp. Sig. (2-tailed)	.969

In another comparison between the mean rank of the writing original sentences with L1 glossing group and control group using post-hoc test of Mann-Whitney U, which can be seen in Table 7 and Table 8, revealed that the retention mean scores of writing original sentences with L1 glossing group (*Md*=5, *n*= 33) and control group (*Md*=2, *n*=33) , *U*=277.50, *z*=-3.486, *p*=.000, *d*=.42 for low

proficient groups were significantly different at the $p<.017$ level with a medium effect size.

Table 7. *Mean Ranks of the Sentence Making with L1 Gloss and Control Groups of Low Proficiency*

	tasks	N	Mean Rank	Sum of Ranks
dimension1	B	33	41.59	1372.50
	C	33	25.41	838.50
	Total	66		

Table 8. Mean Ranks of the Sentence Making with L1 Gloss and Control Groups of Low Proficiency

	test
Mann-Whitney U	277.50
Wilcoxon W	838.50
Z	-3.486
Asymp. Sig. (2-tailed)	.000

[a] Grouping Variable: tasks.

The third comparison which was between the mean ranks of gap-filling with L2 glossing group ($Md=4$, $n=33$) and control group ($Md=2$, $n=33$), $U=263.50$, $z=-3.689$, $p=.000$, $d=.45$, indicated a significant difference at the $p<.017$ level with a medium effect size for the involvement load in these two groups. Table 9 and Table 10 present the findings.

Table 9. *Mean Ranks of the Gap-Filling with L2 Gloss and Control Groups of Low Proficiency*

tasks	N	Mean Rank	Sum of Ranks
A	33	42.02	1386.50
C	33	24.98	824.50
Total	66		

Table 10. *Mean Ranks of the Gap-Filling with L2 Gloss and Control Groups of Low Proficiency*

	tests
Mann-Whitney U	263.50
Wilcoxon W	824.50
Z	-3.689
Asymp. Sig. (2-tailed)	.000

[a] Grouping Variable: tasks.

H02: The task-induced involvement has no effect on the retention of high proficient Iranian EFL students.

Kruskal-Wallis Test revealed that the three groups, χ^2 (2, $n=99$)$=18.30$, $p=.000$ showed a statistically significant difference at the $p<.05$ significance level in the vocabulary retention, as can be seen in Table 11 and Table 12. Therefore, the second null hypothesis was rejected. In other words, there was a statistically significant difference in vocabulary mean scores of high proficient groups who did the three tasks. Table 13 demonstrates that the gap-filling and sentence making groups have the same and higher median ($Md=7$) in comparison with the control group ($Md=3$).

Table 11. *Descriptive Statistics of the Vocabulary Tests of High Proficiency in Terms of Ranks*

tasks	N	Mean Rank
A	33	58.11
B	33	59.14
C	33	32.76
Total	99	

Table 12. *Kruskal Wallis Test of the Vocabulary Tests of High Proficiency*

	tests
Chi-square	18.30
df	2
Asymp. Sig.	.000

[a] Kruskal Wallis Test. [b] Grouping Variable: tasks.

Table 13. *Median of the Three Groups on the Vocabulary Test of High Proficiency*

tasks	N	Mdn
A	33	7.00
B	33	7.00
C	33	3.00
Total	99	4.00

Table 14 and Table 15 present the results of the first Mann-Whitney U Test for the retention mean score of gap-filling with L2 glossing group (Md=7, n=33) and writing original sentences with L1 glossing (Md=7, n=33), U=521.50, z=-.301, p=.764, d=.03 for high proficient groups that failed to reveal a statistically significant difference. The effect size of .03 was very small for involvement load applying Cohen (1988) criteria.

Table14. *Mean Ranks of the Gap-Filling with L2 Gloss and Sentence Making with L1 Gloss Groups of High Proficiency*

	tasks	N	Mean Rank	Sum of Ranks
dimension1	A	33	32.80	1082.50
	B	33	34.20	1128.50
	Total	66		

Table 15. *Mean Ranks of the Gap-Filling with L2 Gloss and Sentence Making with L1 Gloss Groups of High Proficiency*

	tests
Mann-Whitney U	521.50
Wilcoxon W	1082.50
Z	-.301
Asymp. Sig. (2-tailed)	.764
[a] Grouping Variable: tasks.	

In the second comparison between the mean rank of the writing original sentences with L1 glossing group and control group using post-hoc test of Mann-Whitney U, which can be seen in Table 16 and Table 17, revealed that the retention mean scores of writing original sentences with L1 glossing group (*Md*=7, *n*=33) and control group (*Md*=3, *n*=33), *U*=266.00, *z*=-3.631, *p*=.000, *d*=.44 for high proficient groups was significantly different at the p<.017 level for a medium effect size.

Table 16. *Mean Ranks of the Sentence Making L1 Gloss and Control Groups of High Proficiency*

	tasks	N	Mean Rank	Sum of Ranks
dimension1	B	33	41.94	1384.00
	C	33	25.06	827.00
	Total	66		

Table 17. *Mean Ranks of the Sentence Making L1 Gloss and Control Groups of High Proficiency*

	Tests
Mann-Whitney U	266.00
Wilcoxon W	827.00
Z	-3.631
Asymp. Sig. (2-tailed)	.000
[a] Grouping Variable: tasks.	

The third comparison, which was between the mean ranks of the gap-filling with L2 glossing group (*Md*=7, *n*=33) and control group (*Md*=3, *n*=33), *U*=254.00, *z*=-3.780, *p*=.000, *d*=.46, revealed a statistically significant difference at the p< .017 level with a medium effect size for the involvement load in these two groups. Table 18 and Table 19 present the findings.

Table18. *Mean Ranks of the Gap-Filling with L2 Gloss and Control Groups of High Proficiency*

	Tasks	N	Mean Rank	Sum of Ranks
dimension1	A	33	42.30	1396.00
	C	33	24.70	815.00
	Total	66		

Table19. *Mean Ranks of the Gap-Filling with L2 Gloss and Control Groups of High Proficiency*

	Tests
Mann-Whitney U	254.00
Wilcoxon W	815.00
Z	-3.780
Asymp. Sig. (2-tailed)	.000
[a] Grouping Variable: tasks.	

7. Discussion and Conclusions

These research hypotheses examined different tasks in terms of the proficiency of the participants to investigate task-induced involvement. As a result, the null hypotheses were rejected since significant differences were found from administering the tasks. Although some studies (e.g., Keating, 2008; Kim, 2008) were in parallel with this theory's predictions, our study made a more clear-cut

distinction. It was revealed that low proficient participants supported involvement load hypothesis in learning target words through the three tasks. In other words, the results of Dutch-English experiment of Hulstijn and Laufer (2001) and of Kim (2008) supporting the superiority of tasks with higher involvement indices in vocabulary learning were achieved. The sentence making group yielded higher retention than gap-filling and control groups. Besides, the sentence making group did not create significantly higher retention than gap-filling group. On the contrary, high proficient participants behaved the same in gap-filling and sentence making groups but not in control group. Understandably, gap-filling and sentence making groups yielded the same and significantly higher retention than control group.

As it was explained, Laufer and Hulstijin (2001) conducted their research with high proficient participants in 2001 and generalized their hypothesis to all proficiencies. Furthermore, some studies (i.e., Kim, 2008; Keating, 2008) showed that proficiency level does not play a part in task involvement load. In the present study, conversely the involvement load hypothesis only came true for low proficient learners. In other words, vocabulary retention is not only the outcome of deliberate manipulation of task induced involvement (i.e., need, search and evaluation) but also the level of proficiency. Although they knew the role of each component of the involvement load similar, Kim (2008) and the current study proved the other way round. Understandably, it can be stated that devoting the same value to the involvement index should be done in a more conservative method especially with regard to evaluation component. Our findings endorsed the validity of involvement load and in effect the priority of evaluation component over search and need components but only for low proficient learners.

To sum up, the sentence making group which requires deeper cognitive effort (Swain, 1985, 1995) because of inducing learners to compare and make original sentences has the most beneficial influence on vocabulary learning but only for low proficient learners. The evaluation component in high proficient learners is a mysterious construct, which should be more finely graded.

7.1. Suggestions and Implications

Based on the results of the current study, it can be found that tasks which are used in any pedagogical context should be selected considering the language proficiency levels of the learners. Albeit writing original sentences is a challenging task for low intermediate proficient learners, it is of a lot of usability for them. Additionally, as stratification of different tasks is with regard to their level of involvement in low proficient learners, the level of difficulty of word items can come in to scene in teaching and learning vocabularies. In other words, writing original sentences tasks and gap-filling task can be used for teaching difficult words (e.g., technical words) and simpler words, respectively. Another implication concerns with EFL students who have a more restricted time of instruction. They can exploit task involvement to have more time efficacy in vocabulary learning.

7.2. Limitations of the Study

The current investigation like any other research in SLA is liable to some limitations. As time-on-task was not considered in our study, a usual time period was devoted to each task. Longer time on task or longer exposure to the target words can be enumerated as an attribute which can affect the results of word acquisition.

Another limitation of the current investigation is that most of the previous studies confessed the superiority of making sentence over gap-filling groups because of the strong evaluation which these tasks brought about. Furthermore, one feature of these tasks was observing one exposure for each unknown word item. Conversely, in the current study we made learners do a more accurate work (i.e., reading the text, answering the multiple-choice questions, and writing original sentences). Although the participants did not have repeated exposure in the reading text, this factor can be an explanation for word retention too.

Finally, teachers' attitude, students' psychology (Lee, 2003), and type of teachers' reinforcement (Hulstijn & Laufer, 2001) were among the other explanations for higher retention. Since in the present study they were not so much in our control except the teachers' reinforcement, further researches need to be conducted in order to make us sure that the obtained results are due to the task involvement and the participants' level of proficiency.

References

Cheng, H. Ch. (2011). Vocabulary acquisition in learning English as a second language: Examining the involvement load hypothesis and language anxiety with Taiwanese college students (Doctoral dissertation). Retrieved from [http://www.worldcat.org/title/]

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

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

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*, 539-558.

Keating, G. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365-386.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285-325.

Laufer, B. & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22* (1), 1-26.

Lee, C. (2003). *The Effect of Involvement Load on English Vocabulary Acquisition in an Incidental Learning Situation* (Unpublished master's thesis). Hsinchu, Taiwan: National Tsinghua University.

Mackey, A. & Gass, S. M. (2005). *Second language research methodology and design.* Mahwah, N.J.: Lawrence Erlbaum.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York: Newbury House Harper Row.

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

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. H. Hulstijn, & R. Schmidt (Eds.) Consciousness in Second Language Learning. *AILA Review, 11*, 11-26.

Schmidt, R. (2000). Attention. In P. Robinson (Eds.) *Cognition and second language instruction.* Cambridge: Cambridge University Press, pp. 3–32.

Solso, R. L. (1988). *Cognitive psychology* (2nd ed.). Boston, Mass.; London: Allyn and Bacon.

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). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidhofer (Eds.), *Principles and practice in the study of language* (pp. 125-144). Oxford: Oxford University Press.

Walsh, M. I. (2009). *The involvement load hypothesis applied to high school learners in Japan: Measuring the effects of evaluation* (Doctoral dissertation). Retrieved from [http://birmingham.ac.uk/mateflates/]

Xu, J. (2009). An experimental study on the effects of different reading tasks on L2 vocabulary acquisition. *English Language Teaching*, 2, 69-76.

Abbreviations

L1   first language

L2   second language

EFL   English as a foreign language

K-S   Kolmogorov-Smirnov test