The Journal of Teaching Language Skills (JTLS)
5 (2), Summer 2013, Ser. 71/4
ISSN: 2008-8191. pp. 1-26

The Effect of Four Different Types of Involvement Indices on Vocabulary Learning and Retention of EFL Learners

S. Baleghizadeh *
Associate Professor, TEFL
Shahid Beheshti University, Tehran
email: sasanbaleghizadeh@yahoo.com

M. Abbasi
M.A., TEFL
Shahid Beheshti University, Tehran
email: maryam.abbasi55@gmail.com

Abstract

The purpose of the present study was to provide empirical support for the construct of the involvement load hypothesis (ILH) in an EFL context. To fulfill the purpose of the study, 4 intact groups consisting of 126 intermediate-level students participated in this experiment. In order to ensure that the participants were at the same level of English language proficiency, the Nelson test was administered prior to the treatment. Moreover, the participants were pretested on the knowledge of the target items through the Vocabulary Knowledge Scale (VKS). During the 7 treatment sessions, the 4 groups were treated with different tasks (reading, fill-in-the-blanks, sentence-writing, and composition-writing) varying in the involvement index according to the ILH. The VKS was administered twice (immediate and delayed posttests) to measure the gain degree at receptive and productive levels. The results indicated the validity of the hypothesis in receptive and productive learning and receptive retention. In productive retention, however, partial support for the hypothesis was provided. In addition, vocabulary gain in partially known, receptive, and productive categories could lend support to the effectiveness of each treatment over time.

Keywords: EFL, vocabulary, involvement load hypothesis (ILH), the involvement index

Received: 1/5/2013 Accepted: 5/21/2013

* Corresponding author

1. Introduction

In order to put an end to the "model-free" (Meara, 1997, p. 111) nature of vocabulary learning, Laufer and Hulstijn (2001) proposed the involvement load hypothesis. In this model, attempts are made to draw on cognitive (elaboration, attention as well as implicit, and explicit learning) and affective (motivation and need) aspects of L2 learning. Moreover, the framework was proposed in a situation that levels of processing (LOP) by Craik and Lockhart (1972) had opened another gap in the literature. The LOP favored to talk of three levels of perceptual processing while considering memory: physical or sensory analysis, pattern recognition, and stimulus elaboration and enrichment. Craik and Lockhart (1972) consider the last level to play the most significant role in the long-term retention of the items presented because it is in charge of manipulation, more elaboration, and deeper processing of those items. A number of scholars, however, (Braddeley, 1978; Eysenck, 1978; Nelson, 1977) have cast doubt on the validity of this framework. Their major concern is with indexing the depth. Nevertheless, the aforementioned model is unable to provide this index. In order to "operationalize [the] general cognitive notions" (Hulstijn & Laufer, 2001, p. 543) introduced by this outdated framework, the involvement load hypothesis (ILH) was developed. This model considers three different dimensions of each vocabulary task: need, search, and evaluation.

According to Laufer and Hulstijn (2001, p.14), need is the motivational dimension of the model, which is concerned with the achievement desire. This notion stands for "a drive to comply with the task requirements" (p. 14). Hulstijn and Laufer (2001) make a distinction between two types of needs, namely "moderate and strong" (p. 543) to refer to different degrees of this drive. Moderate need is imposed on the learner by an "external agent" (p. 543) and gives an index of one. As an example, if a reading task contains an unfamiliar word that is a prerequisite for understanding the passage, the learner will experience the need to figure it out. On the other hand, strong need is "self-imposed" (p. 543) and induces a load of two. An example is when the learner wants to refer to an object or concept and the L2 form is unfamiliar (Laufer & Hulstijn, 2001).

According to Laufer and Hulstijn (2001), "Search and evaluation are the two cognitive (information processing) dimensions of involvement" (p. 14). Search refers to the attempt of the learner to find a meaning for an unknown L2 word or the L2 word form to express the required concept by referring to a dictionary or another authority such as a teacher (Hulstijn & Laufer, 2001). This component of the construct can be either absent (index = 0) or present (index = 1).

Evaluation, as the last constituent, refers to the decision to be made based on "semantic and formal appropriateness (fit) of the word and its context" (Laufer & Hulstijn, 2001, p. 15). This component can be either moderate or strong. The former refers to the occasions when "differences between words (as in a fill-in task with words provided) or differences between several senses of a word in a given context" (p. 15) are required. An example is a reading task, during which a word is looked up and a homonym is found. Hence, the learner has to compare all the possible meanings against the given context and choose the most appropriate one (Hulstijn, 1992; Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001). This component induces an index of one. On the contrary, the type of evaluation that requires "a decision about additional words which will combine with the new word in an original sentence or text" (Laufer & Hulstijn, p. 15) is referred to as strong evaluation (index = 2). As an example to illustrate this case, an L2 writing task can be considered. In order to translate back the L1 ideas into an L2 piece of writing, the target items must be looked up in a dictionary. If the word has more than one translation, "additional syntagmatic decisions about the precise collocations" of that item have to be made (p. 15). The absence of this component, however, carries no load (index = 0). In order to establish the effectiveness of each vocabulary task, the index induced by the three hypothetical components (need, search, and evaluation) is counted. Consequently, the higher induced index is expected to result in the better retention of the items (Laufer & Hulstijn).

2. Literature Review

In order to provide experimental support for the construct of the ILH, Hulstijn and Laufer (2001) conducted two parallel experiments on advanced Dutch and Hebrew learners. They had three groups in each study for whom they provided different tasks. The first group was a reading condition, in which 10 target words were highlighted and glossed (L1). In addition, the participants were expected to answer the comprehension questions (10 multiple-choice) after the main task. As a result, this condition called for moderate need (1), no search (0) and no evaluation (0). Overall, the involvement index was 1 (1+0+0). In the second group, the participants were provided with the same passage. However, the 10 items were replaced with blanks to fill in. The target words along with five distractors were glossed (L1 translations and L2 definitions) on a separate sheet. This task entailed an index of 2 (1+0+1). The third condition required the participants to create a composition containing the 10 items. The teacher, in this task, was responsible for explaining the words' meanings and giving examples of the words' use. Therefore, this task required moderate need (1), no search (0)

and strong evaluation (2). As a result, the involvement index was 3 (1+0+2). According to what Hulstijn and Laufer suggested, the composition group was expected to reveal better performance in comparison with the other two in both tests (immediate and delayed). They also hypothesized that the fill-in group would yield better results than the reading condition, due to the higher load it induced. Consequently, the results obtained from the Dutch participants did confirm the first part of their hypothesis. However, the second part could not be proven. The results of their corresponding study with Hebrew participants, however, fully supported their hypothesis.

In another study investigating the ILH, Keating (2008) was concerned with the degree of passive/active word learning and retention of beginning Spanish learners. His participants in each group completed a different task with varying indices: reading comprehension (1), reading plus word suppliance (2), and sentence-writing (3). Considering the ILH, the third task, which induced the highest involvement, was expected to be more beneficial. The conclusions he drew in the immediate posttest of active recall were in line with the hypothesis. In the delayed active recall, although the mean score of task 2 was significantly greater than 1, the third task was no more superior to 1 and 2. In the passive recall posttests, although the second and third tasks were more effective than the first one, the third one could not establish its superiority to the second. Moreover, considering the time on task, Keating suggests that the benefits associated with more demanding tasks discolor.

Jing and Jianbin (2009) tested the hypothesis to confirm vocabulary learning and retention in the listening comprehension tasks. In task A, the participants were provided with marginal glosses of the new words and were expected to answer comprehension questions. However, the participants were able to answer them without any understanding of the items. The load of this task was 0 (0+0+0). The participants in task B, however, needed the understanding of the items to answer the questions (1+0+0=1). Task C required the participants to do the same job of group B along with writing a short article with the target words. Therefore, this task induced an index of 3 (1+0+2). Subsequent scores of the immediate and delayed posttests were able to shed light on the new hypothesis.

In another study carried out by Nasrollahy Shahry (2010), two types of tasks, namely reading comprehension and sentence-writing across two proficiency levels (low-intermediate and high-intermediate) were examined. In the reading condition, the participants read sentences containing the target words, but in the writing condition, they wrote sentences with those items. Learning and retention of the items were measured twice (immediate and delayed) and at two levels (receptive and productive). The results

demonstrated an advantage for the lower-proficiency writing over the reading group in both posttests which was congruent with the hypothesis. In the high proficiency group, however, no significant differences were observed.

Kim (2011) carried out two experiments in order to examine the construct of the ILH. The design of her first study was similar to the one conducted by Hulstijn and Laufer (2001). However, she carried out her research along two proficiency levels: undergraduate and intensive English program students. She obtained similar results in both levels: In the immediate and delayed posttests, the composition group could outperform the fill-in and the reading groups. Furthermore, in the delayed posttest, the fill-in group was able to outperform the reading condition. In the immediate posttest, however, no significant difference was observed between the fill-in and the reading groups. In her second study, Kim had four groups of participants with two proficiency levels to which she randomly assigned a different task (composition or sentence-writing). Considering the involvement load, these two tasks induced the same index (1+0+2=3). As a result, no significant differences were observed between their performances.

Regarding the significance of the ILH, it seems essential to have more experiments, considering its specific features. In each of the earlier experiments (Hulstijn & Laufer, 2001; Jing & Jianbin, 2009; Keating, 2008; Kim, 2011; Nasrollahy Shahry, 2010), three types of tasks were dealt with, leaving out other possible conditions. Thus, the present study is concerned with filling this gap and examining the four types of loads in a single study. In addition, this experiment departs from the literature in its duration over seven sessions which might add to its ecological validity.

3. Research Questions

To investigate the effectiveness of the ILH, the following research questions guided the study:

1) Is there any significant difference among the reading (R), filling-in-the-blanks (FIB), sentence-writing (SW) and composition-writing (CW) groups concerning the receptive and productive learning of the target items?
2) Is there any significant difference among the treatment groups (R, FIB, SW, and CW) concerning the receptive and productive retention of the target items?
3) Is there any significant difference among the groups (R, FIB, SW, and CW) from the pretest to the immediate and delayed posttests regarding the receptive and productive aspects?

4) Is there any difference among the groups (R, FIB, SW, and CW) from the pretest to the immediate and delayed posttests regarding the percentages of the unknown, partially known, receptive, and productive categories?

4. Method

4.1 Participants

The present study was carried out in a private language institute in Hamedan, Iran, involving 126 EFL learners (53 males and 73 females) ranging in age from 11 to 13 as the participants. 32 Thirty-two participants took part in each of the R and CW groups and 31 in each of the other two (FIB and SW). At the time of the study, the proficiency level of the participants was intermediate and the *Interchange* series were the materials they were studying. The original pool included 128 learners, from whom the data were obtained. However, the data from two of the participants had to be discarded because they missed one of the posttests.

4.2 Target words

In the first treatment session, 10 items and in the remaining six, 12 words were the focus of attention, and overall the participants were provided with 82 words (from the four important syntactic categories: nouns, verbs, adjectives, and adverbs). An important concern of the researchers for choosing the items was their unfamiliarity to the participants. Consequently, lower frequent words were selected from the *504 Absolutely Essential Words* (Bromberg, Liebb, & Traiger, 1996) and an official Web site for reading instruction (www.ReadTheory.Org2010). The target items of the study and their frequency are illustrated in Table 1. (www.collinsdictionary.com):

Table 1. The target items of the study and their commonness

Commonness*	Target Items
☆☆☆☆☆★	vaporize
☆☆☆☆★★	pierce – swarm – kneel – drenched – tumult – wobble – blizzard – cardigan – sled – shovel – flashlight – shriek – frigid – numb – wager – pastry – outlaw – captivate – light up – pitcher – ogre – reverent – sheriff – shove
☆☆☆★★	surgeon – slender – cautious – retina – surpass – charity – tyrant – furthermore – rebel – blast – figure out – infrequently – moist – splendid – obstacle – converge – conceal – corpse – hedge – menacingly – tempt – bait – decay – deceive – disclose – nutrition – bunch – deteriorate – lucid – monotonous – reckless – revenge –torture – threaten – unequivocally – transgression

Commonness*	Target Items
☆☆☆☆☆	beam – eventually – hostility – invasion – tightly – folk – abandon – grave – disaster – cease – excessive – promote – survey – conveniently – delicate – magic – stretch – subsequently – persuade – shelter – efficacy
★★★★★	---

*The number of stars indicates the frequency of the target items.

4.3 Tasks

The participants in each group (R, FIB, SW, and CW) were provided with different tasks, which resembled different treatment conditions considering the induced involvement index (see Appendices B, C, D, and E). In the R condition, the participants were presented with a short passage (165 words on average) and some comprehension questions to answer (true/false and multiple-choice). The target words appeared in bold to ensure that the participants would notice them with no difficulty (Gass, 1988; Paribakht & Wesche, 1997). The passages were taken from the same sources as the target items along with some modifications. These alterations were aimed at dealing "with difficult words and phrases by replacing them with more familiar synonyms or paraphrases" (Read, 2000, p. 194) in order to increase the readability of the text.

For the FIB group, these items were presented above the incomplete sentences along with a distractor. Therefore, the participants of this group had to be presented with 13 items in the definition sheet. Otherwise, they could have easily recognized the distractor as the item not provided there. The researchers made an attempt to choose the distractors from the part of speech that was in minority to minimize the effect of grammatical guesswork. In addition, the *Longman Dictionary of Contemporary English* (2009), the *Cambridge Advanced Learner's Dictionary* (2008) and the *Oxford Advanced Learner's Dictionary* (2005) were the sources the gapped sentences were obtained from.

The participants of the SW group had to write sentences with each of the items. Although they could use the example sentences on the definition sheet as a guide, they were expected to write an original one.

Finally, the participants in the CW condition had to write a short paragraph containing the new words or two very short paragraphs, in case they could not organize their imagination in one. The topics of the reading condition were provided for this group as an optional hint, yet they did not have to use them.

In addition, in each session of the study, the participants were provided with a sheet to make their comprehension of the main task easier. Containing L1 translations (Persian), L2 definitions (English), and a sample

sentence illustrating the new words, it was called the definition sheet. The sentences were taken from the three aforementioned dictionaries with the items organized alphabetically to give the participants a sense of using a very small dictionary (Folse, 2006; Nasrollahy Shahry, 2010). Furthermore, the participants were provided with the target words of each session separately, in order not to have pre-exposure to the other treatment items (Nasrollahy Shahry, 2010).

With regard to the index load carried by each task, in all the aforementioned conditions, moderate need was induced (1), as the tasks were teacher-imposed. Similarly, no search was carried (0) because the definitions and translations were provided for the participants. However, in each condition, a different degree of the evaluation was marked. In the R condition, the absence of this component induced a zero load. Overall, the index load for such task was 1 (1+0+0). The FIB condition required the participants to make difference between the target items; as a result, moderate evaluation was carried (1+0+1). The SW and CW conditions, however, induced a higher index because combination of the old and new items was taken into account (index load = 3, 1+0+2).

In order to establish the appropriateness of the tasks, their running words, the example sentences, and definitions for instructional purposes, the views of two experienced teachers of the same institute were elicited. This elicitation procedure for the reading group was also in the form of ranking the tasks from the easiest to the more difficult ones.

4.4 Pretest, immediate, and delayed posttests

The participants received a pretest at the beginning of the treatment and two posttests at the end. The original format of the Vocabulary Knowledge Scale (VKS), developed by Paribakht and Wesche (1993, 1997), was the basis for such assessment:

Table 2. The vocabulary knowledge scale (Paribakht & Wesche, 1997, p. 180)

I. I don't remember having seen this word before.
II. I have seen this word before, but I don't know what it means.
III. I think I know the meaning of this word. It means _____ (synonym or translation)
IV. I know this word. It means _____ (synonym or translation)
V. I can use this word in a sentence:
_____ (Write a sentence.)
(If you do this section, please also do section IV.)

For scoring the responses, however, the modified version developed by Min (2008) was taken into account. The reason for such a decision is that VKS scoring in its original format does not allow for quantitative estimates of vocabulary gain. The scoring procedure in the present study is provided below:

- *Unknown*: No points were assigned to this category (I), since zero knowledge was reported in this respect.
- *Partially known*: Zero point was also assigned to this category (II), since the participants reported no knowledge of the item by choosing it. However, they demonstrated partial knowledge of the form. In addition, if the participants produced the wrong synonym or translation for the target word in category III, this was also considered as partial knowledge of the form.
- *Receptive*: The participants got a score of one if they were able to provide the correct synonym or translation for the target item (IV). The participants were not penalized for minor spelling mistakes in writing the synonyms.
- *Productive*: A score of one was assigned to semantically appropriate sentences (V) without considering minor spelling mistakes.

5. Procedure

Before carrying out the treatment (first session of the semester), the Nelson English Language Test at 250C level (Fowler & Coe, 1976) was administered to ensure the participants' homogeneity with regard to their general English language proficiency. In the second session, the VKS was given to the participants as a measure of their pre-exposure to the target items. Consulting with the previous teachers of the participants, the researchers assumed that they would know approximately none of the items.

The treatment started from the third session and extended over seven sessions. During these sessions, the four groups received different treatments with different loads. However, something that was provided for all the participants in the same way was the definition sheet. The teacher, for the four groups, pronounced the vocabulary items on this sheet and provided explanations if there were problems. For the R group, the feedback was mostly in the form of checking the answers to the questions to make sure the participants had understood the passage and the new items. The same procedure was followed for the FIB condition: checking the words participants had put in the blank spaces. For the SW and CW groups, the teacher walked around the classroom checking the written product of the participants and helping them where it deemed necessary.

An important issue to consider was the notion of time on task, since it is believed "that task effectiveness is a function of time spent on task" (Keating, 2008, p. 379). However, Hulstijn and Laufer (2001) consider it "as an inherent property of a task, not as a separate variable" (p. 549). In order to address this probable concern, the researchers pilot-tested three of the tasks of each group randomly to classes of the same level to have a rough estimate of the time spent on each task. The R, FIB, SW and CW tasks took 19, 21, 30, and 33 min on average to be completed respectively. As a result, we decided to give the time spent on the most time-consuming task (CW) to the other groups. However, they were free to submit their handouts whenever they finished the task.

In the tenth session, an immediate posttest in the form of the VKS was administered to the participants, which provided the opportunity for the researchers to reflect on the participants' gain from receptive and productive perspectives. One month after the administration of the first posttest, the participants took a delayed posttest (VKS) to indicate their retention degree of the items.

6. Results

6.1 Nelson English language test

For the purpose of homogeneity establishment, the Nelson test with a maximum score of 50 was administered before the treatment. Subsequently, the ANOVA results indicated that there was no significant difference among the groups at this time point (F = .128, df = 3) because the p value was above the alpha level (p = .943).

6.2 Receptive and productive pretest scores

For the receptive aspect, approximately no knowledge of the items was reported because only two of the participants (one in each of the FIB and SW conditions) could provide L1 translations for one of the items (out of 82). Similarly, for the productive dimension, zero knowledge was indicated. Thus, the unfamiliarity of the participants with the target items before the beginning of the treatment was ascertained.

6.3 Receptive and productive immediate posttest scores

In order to measure the degree of receptive and productive gains immediately after the treatment, the VKS was administered. Similar to the pretest, the results of this test are reported out of 82 displayed in Table 3:

Table 3. Descriptive statistics of the receptive and productive immediate posttest

	Groups	*N*	Mean	Std. Deviation	Minimum	Maximum
Receptive	**R**	32	22.96	3.26	16.00	27.00
	FIB	31	25.96	3.74	19.00	30.00
	SW	31	29.03	3.48	23.00	33.00
	CW	32	30.09	3.50	25.00	35.00
	Total	126	27.00	4.45	16.00	35.00

	Groups	*N*	Mean	Std. Deviation	Minimum	Maximum
Productive	**R**	32	15.96	2.59	10.00	19.00
	FIB	31	19.03	3.22	10.00	21.00
	SW	31	24.96	3.68	20.00	30.00
	CW	32	25.03	3.05	21.00	32.00
	Total	126	21.23	5.01	10.00	32.00

The statistical data analyses were based on an ANOVA with the treatment type as a between groups variable. Furthermore, for all the analyses the alpha level was set at .05. Table 4 illustrates the results of such analyses:

Table 4. ANOVA results of the receptive and productive immediate posttest

Receptive	Sum of Squares	*df*	Mean Square	*F*	*Sig.*
Between Groups	987.36	3	329.12	26.81	.000
Within Groups	1497.62	122	12.27		
Total	2484.99	125			

Productive	Sum of Squares	*df*	Mean Square	*F*	*Sig.*
Between Groups	1930.98	3	643.66	64.47	.000
Within Groups	1217.87	122	9.98		
Total	3148.85	125			

The results in Table 4 indicate that there was a significant difference among the groups at this point of time (receptive: $F = 26.81$, $df = 3$; productive: $F = 64.47$, $df = 3$) because the p value was below the alpha level ($p = .000$ in both cases). Therefore, the first research question was answered in the

positive. In order to show where this significant difference had occurred, the post hoc Scheffe was run, the results of which are represented in Table 5

Table 5. Pair wise comparisons of the receptive and productive immediate post-test

Paired Tests (Receptive)	Mean Difference	Std. Error	*Sig.*
R/ FIB	-2.99*	.88	.011
R/ SW	-6.06*	.88	.000
R/ CW	-7.12*	.87	.000
FIB/ SW	-3.06*	.88	.01
FIB/ CW	-4.12*	.88	.000
SW/ CW	-1.06	.88	.696

Paired Tests (Productive)	Mean Difference	Std. Error	*Sig.*
R/ FIB	-3.06*	.79	.003
R/ SW	-8.99*	.79	.000
R/ CW	-9.06*	.78	.000
FIB/ SW	-5.93*	.80	.000
FIB/ CW	-5.99*	.79	.000
SW/ CW	-.063	.79	1.000

*The negative mean difference indicates the mean of the first group is lower than that of the second group.

As shown in Table 5, the mean difference between the FIB and R groups (receptive: *MD* = -2.99; productive: *MD* = -3.06) indicates that the FIB participants made significant vocabulary gains after the treatment period (receptive: p = .011; productive: p = .003). The same results hold for the superiority of the SW and CW groups to the R condition because the mean differences (receptive: *MD* = -6.06, *MD* = -7.12; productive: *MD* = -8.99, *MD* = -9.06, respectively) were significantly higher in all the cases (p = .000). In addition, the FIB condition turned out to be less effective than the SW or CW treatments because the mean difference (receptive: *MD* = -3.06, *MD* = -4.12; productive: *MD* = -5.93, *MD* = -5.99) was lower in each case. As a result, the participants in the SW and CW groups made significant gains (receptive: p = .01, p = .000; productive: p = .000, p = .000) in this regard. Regarding the last two conditions (SW and CW), however, the mean difference (receptive: *MD* = -1.06; productive: *MD* = -.063) did not lead to any significant differences (receptive: p = .696; productive: p = 1.000).

6.4 Receptive and productive delayed posttest scores

The VKS was administered one month after the first posttest with the aim of measuring the degree of retention. Similar to the pretest and immediate posttest, the results of this test are reported out of 82. Table 6 presents the descriptive statistics of these scores:

Table 6. Descriptive statistics of the receptive and productive delayed posttest

	Groups	*N*	Mean	Std. Deviation	Minimum	Maximum
Receptive	**R**	32	19.06	3.03	11.00	22.00
	FIB	31	21.83	3.33	16.00	26.00
	SW	31	25.03	3.45	19.00	29.00
	CW	32	25.90	3.74	20.00	32.00
	Total	126	22.95	4.33	11.00	32.00

	Groups	*N*	Mean	Std. Deviation	Minimum	Maximum
Productive	**R**	32	13.03	2.42	8.00	16.00
	FIB	31	15.29	3.27	10.00	20.00
	SW	31	19.12	3.60	13.00	23.00
	CW	32	20.09	3.43	12.00	24.00
	Total	126	16.88	4.29	8.00	24.00

The data analysis was based on an ANOVA with the treatment type as a between groups variable. Table 7 illustrates the results of such analysis:

Table 7. ANOVA results of the receptive and productive delayed posttest

Receptive	Sum of Squares	*df*	Mean Square	*F*	*Sig.*
Between Groups	935.95	3	311.98	26.92	.000
Within Groups	1413.75	122	11.58		
Total	2349.71	125			

Productive	Sum of Squares	*df*	Mean Square	*F*	*Sig.*
Between Groups	1039.65	3	346.55	33.51	.000
Within Groups	1261.55	122	10.34		
Total	2301.21	125			

As indicated in Table 7, a significant difference among the groups regarding their receptive (F = 26.92, df = 3) and productive (F = 33.51, df = 3) retention was observed because the p value was below the alpha level (p = .000 in both cases). As a result, the second research question was answered

in the affirmative. In order to spot the place where these differences had occurred, Table 8 displays the results of post hoc Scheffe:

Table 8. Pair-wise comparisons of the receptive and productive delayed posttest

Paired Tests (Receptive)	Mean Difference	Std. Error	*Sig.*
R/ FIB	-2.77*	.85	.01
R/ SW	-5.96*	.85	.000
R/ CW	-6.84*	.85	.000
FIB/ SW	-3.19*	.86	.005
FIB/ CW	-4.06*	.85	.000
SW/ CW	-.87	.85	.792

Paired Tests (Productive)	Mean Difference	Std. Error	*Sig.*
R/ FIB	-2.25	.81	.056
R/ SW	-6.09*	.81	.000
R/ CW	-7.06*	.80	.000
FIB/ SW	-3.83*	.81	.000
FIB/ CW	-4.80*	.81	.000
SW/ CW	-.96	.81	.702

As suggested by Table 8, the FIB group could outperform the R group because the receptive mean difference (*MD* = -2.77) was significantly superior (p = .01). However, this picture changed in the productive condition, as the mean difference (*MD* = -2.25) could not reach the significance level (p = .056). In addition, the SW and CW groups revealed a higher receptive mean difference (*MD* = -5.96, *MD* = -6.84) in comparison with the R condition that reached the significance level (p = .000 in both cases). The same holds for the productive retention, since the participants of the SW and CW treatments were able to outperform those of the R group, owing to the high mean difference observed (*MD* = -6.09, *MD* = -7.06) that could reach the significance level (p = .000 in both cases). Furthermore, the SW and CW conditions could also outperform the FIB group because the mean difference in receptive (*MD* = -3.19, *MD* = -4.06) and productive aspects (*MD* = -3.83, *MD* = -4.80) was observed as being significant (receptive: p = .005, p = .000; productive: p = .000, p = .000). Regarding the difference between the last two groups, however, the receptive (*MD* = -.87) and productive mean differences (*MD* = -.96) could not display any

significance difference in both cases (receptive: p = .792; productive: p= .702).

6.5 Repeated measures ANOVA for the receptive and productive scores

The descriptive statistics of the pretest, immediate and delayed post-tests of the four conditions are reported in Table 9. The immediate posttest scores were considered as learning scores, whereas those of the delayed test as retention scores:

Table 9. Descriptive statistics for the receptive and productive scores measured over time

Group (Receptive)	Test	Mean	Std. Deviation	*N*
R	Pretest	.00	.00	32
	Posttest 1	22.96	3.26	32
	Posttest 2	19.06	3.03	32
FIB	Pretest	.032	.17	31
	Posttest 1	25.96	3.74	31
	Posttest 2	21.83	3.33	31
SW	Pretest	.032	.17	31
	Posttest 1	29.03	3.48	31
	Posttest 2	25.03	3.45	31
CW	Pretest	.00	.00	32
	Posttest 1	30.09	3.50	32
	Posttest 2	25.90	3.74	32

Group (Productive)	Test	Mean	Std. Deviation	*N*
R	Pretest	.00	.00	32
	Posttest 1	15.96	2.59	32
	Posttest 2	13.03	2.42	32
FIB	Pretest	.00	.00	31
	Posttest 1	19.03	3.22	31
	Posttest 2	15.29	3.27	31
SW	Pretest	.00	.00	31
	Posttest 1	24.96	3.68	31
	Posttest 2	19.12	3.60	31
CW	Pretest	.00	.00	32
	Posttest 1	25.03	3.05	32
	Posttest 2	20.09	3.43	32

A repeated measures analysis of variance (ANOVAR) was incorporated in order to answer the third research question. The assessment time (pretest, immediate, and delayed posttests) was considered as a within-subjects variable and instructional treatment (R, FIB, SW, and CW) as a between-subjects variable. For all the statistical analyses, the alpha level was set at .05:

Table 10. Repeated measures ANOVA for the receptive and productive scores measured over time

Source (Receptive)	SS	*df*	*F*	*P*	*η*2
		Between-subjects			
Intercept	104938.62	1	9833.28	.000	.98
Treatment	1283.01	3	40.07	.000	.49
Error	1301.95	122			
		Within-subjects			
Time	53403.79	2	4043.34	.000	.97
Time/Treatment	640.34	۶	16.16	.000	.28
Error	1611.35	244			

Source (Productive)	SS	*df*	*F*	*P*	*η*2
		Between-subjects			
Intercept	61067.78	1	8268.93	.000	.98
Treatment	1928.74	3	87.05	.000	.68
Error	900.99	122			
		Within-subjects			
Time	31733.34	2	2452.72	.000	.95
Time/Treatment	1041.89	۶	26.84	.000	.39
Error	1578.43	244			

The ANOVAR results indicated that the amount of gain varied as a function of treatment (receptive: $F = 40.07$, $df = 3$, error $df = 122$; productive: $F = 87.05$, $df = 3$, error $df = 122$) and time of measurement (receptive: $F = 4043.34$, $df = 2$, error $df = 244$; productive: $F = 2452.72$, $df = 2$, error $df = 244$). The effect sizes ($\eta 2$) for the treatment and time were .49 and .97 (receptive) and .68 and .95 (productive), respectively, which showed each factor was able to account for a substantial variance in the scores. Moreover, the interaction of time and treatment was regarded as

significant (receptive: F = 16.16, df = 6; productive: F = 26.84, df = 6) indicating that different treatments led to different amounts of gain at different time points.

Regarding the effectiveness of each treatment, however, multiple comparisons over time seemed necessary. For the receptive and productive gains of the R group, the Wilks' Lambda measure revealed a significant trend (receptive: F = 954.13, df = 2, p = .000, $\eta 2$ = .985; productive: F = 891.54, df = 2, p = .000, $\eta 2$ = .983). The same trend of significant difference was also observed for the FIB (receptive: F = 1183.17, df = 2, p = .000, $\eta 2$ = .988; productive: F = 1121.71, df = 2, p = .000, $\eta 2$ = .987), SW (receptive: F = 1270.06, df = 2, p = .000, $\eta 2$ = .989; productive: F = 1084.107 , df = 2, p = .000, $\eta 2$ = .987) and the CW conditions (receptive: F = 1312.55, df = 2, p = .000, $\eta 2$ = .989; productive: F = 1124.71, df = 2, p = .000, $\eta 2$ = .987). Therefore, Table 11 demonstrates the results of the post hoc comparison of scores among different time points within each group:

Table 11. Comparisons of scores between paired tests for the four groups

(Receptive)	Paired Tests	Mean Difference	Std. Error	*Sig.*
R	**1/2**	-22.96*	.57	.000
	1/3	-19.06*	.53	.000
	2/3	3.90*	.58	.000
FIB	**1/2**	-25.93*	.67	.000
	1/3	-21.80*	.60	.000
	2/3	4.12*	.84	.000
SW	**1/2**	-29.00*	.63	.000
	1/3	-25.00*	.63	.000
	2/3	4.00*	.69	.000
CW	**1/2**	-30.09*	.62	.000
	1/3	-25.90*	.66	.000
	2/3	4.18*	.66	.000

(Productive)	Paired Tests	Mean Difference	Std. Error	*Sig.*
R	**1/2**	-15.96*	.45	.000
	1/3	-13.03*	.42	.000
	2/3	2.93*	.57	.000
FIB	**1/2**	-19.03*	.58	.000
	1/3	-15.29*	.58	.000
	2/3	3.74*	.92	.001
	1/2	-24.96*	.66	.000

SW	**1/3**	-19.12*	.64	.000
	2/3	5.83*	.91	.000
	1/2	-25.03*	.54	.000
CW	**1/3**	-20.09*	.60	.000
	2/3	4.93*	.59	.000

1: Pre-test 2: Post-test 1 3: Post-test 2

The mean difference (see Table 11) between pretest and posttest 1 of each group signifies that the participants made significant receptive and productive gains at this time point. Furthermore, a significant degree of knowledge retention was indicated in all the conditions, due to the mean difference between pretest and posttest 2. However, a significant degree of knowledge was lost in the gap between the two posttests. Thus, the third research question was answered in the positive.

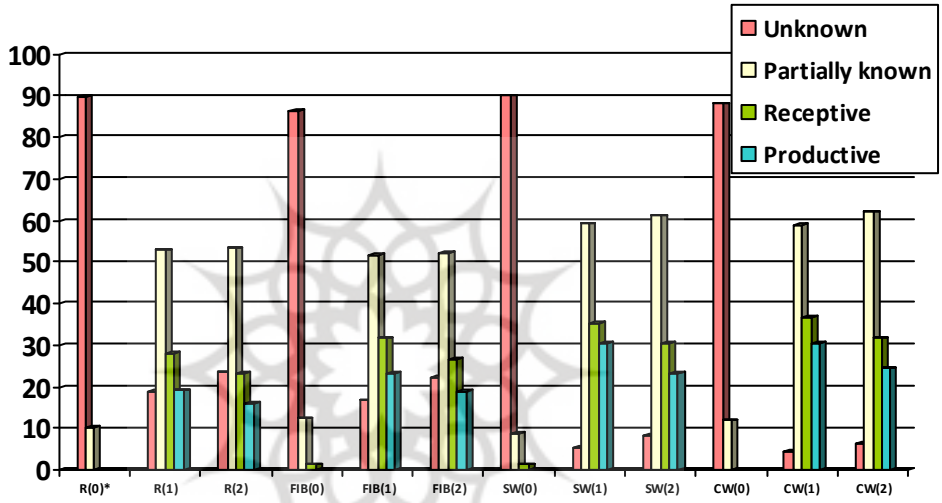
7. Qualitative Changes

Table 12 displays the qualitative changes of the four conditions over time based on the VKS. Thus, the results of the unknown, partially known, receptive, and productive knowledge are presented separately (Min, 2008; Nasrollahy Shahry, 2010). In Table 12, the percentages exceed 100 because the receptive knowledge is too assumed to include productive knowledge. This interrelationship was previously referred to by Paribakht and Wesche (1997), who devised the aforementioned scale.

Table 12. Percentages of unknown, partially known, receptive, and productive vocabulary over time

Group	Test	Unknown	Partially Known	Receptive	Productive
R	Pre-test	89.78	10.22	0	0
	Post-test 1	19	53	28	19.13
	Post-test 2	23.38	53.38	23.24	15.89
FIB	Pre-test	۸۶,۴۲	۱۲,۳۷	1.21	۰
	Post-test 1	16.73	51.62	31.65	23.2
	Post-test 2	22.21	52.17	26.62	18.84
SW	Pre-test	90.01	8.78	1.21	0
	Post-test 1	5.37	59.23	35.4	30.43
	Post-test 2	8.36	61.12	30.52	23.31
CW	Pre-test	88.08	11.92	0	0
	Post-test 1	4.38	58.93	36.69	30.52
	Post-test 2	6.33	62.09	31.58	24.5

As displayed in Table 12, the participants progressed from the unknown category to the partially known, receptive, and productive knowledge. Moreover, all the groups demonstrated a trend of loss from the immediate to the delayed test (Min, 2008). However, the percentage of unknown words in the delayed test was still lower than that reported for the pretest. Figure 1 is a graphical representation of the knowledge gain for the treatment groups over time:



*1: Pre-test 2: Post-test 3: Post-test
1 2

Figure 1 Proportion of Unknown, Partially Known, Receptive, and Productive Vocabulary in the Pretest, Immediate, and Delayed Posttests

In order to answer the fourth research question, a qualitative comparison of the groups' achievement over time seemed necessary. Although all the conditions were successful in decreasing the degree of unknown words and adding to the partially known, receptive and productive categories, the CW and SW treatments—with a slight superiority of the CW— turned out to be the most effective tasks in this regard.

8. Discussion

The results of the receptive and productive learning and receptive retention are in line with what Hulstijn and Laufer (2001) reported for their Hebrew learners. Keating (2008) in his immediate posttest of active recall, Jing and Jianbin (2009), Nasrollahy Shahry (2010) for his lower-proficiency

participants and Kim (2011) in her first experiment (delayed post-test) also reported the same results. Additionally, the non-significant difference between the SW and CW groups corroborated those of Kim's (2011) second experiment. However, the results are inconsistent with what Barcroft (2004, 2006) has suggested about sentence-writing. A possible cause for this sharp difference might be the testing instrument because in his studies the participants were provided with the picture of the target items and had to remember the L2 form of them.

A probable cause for the different performance of the groups might be the different degree of the evaluation induced by each of the tasks. The degree of the need component was equal for all the conditions because the tasks were teacher-imposed. The same was true for the search constituent, as the definitions and translations were provided for all the participants in the definition sheet. However, for the R condition, no evaluation was marked because the participants were not required to make semantic decisions about the target items and their context (Laufer & Hulstijn, 2001). In the FIB condition, a moderate type of evaluation was induced, as the participants had to make differences between the target items to choose the appropriate ones fitting the blanks. For the last two conditions (SW and CW), the highest evaluation was considered because "a decision about additional words which will combine with the new word in an original sentence or text" (p. 15) was taken into account. This pushing force was also referred to by Laufer (1998) as a determining factor in vocabulary studies. This idea corroborated the output hypothesis (Swain, 1985) because the production of language items seems a promising way in enriching and enhancing the input (Swain & Lapkin, 1995).

The superior performance of the participants in these conditions can also lead to the conclusion that they were more conducive to noticing (Schmidt, 1990, 1995, 2000) of the target items. Even if such attention is not taken into account, task demands might be able to make the material be processed at a higher degree (Anderson, 1985; Joe, 1995). Accordingly, it seems that the demands the SW and CW tasks put on the participants were stronger than those put forward by their counterparts.

Regarding the LOP, it can be hypothesized that the SW and CW conditions were able to engage the participants in deeper levels than the R and FIB. The participants of the former groups might have been able to go through the three levels of perceptual processing suggested by Craik and Lockhart (1972), "sensory analyses, pattern recognition, and stimulus elaboration" (p. 676). The last stage in their model seemed to account for the significant difference among the treatment groups that led to "trace persistence" (p. 675)

The superiority of these two groups can also be looked at from another view point: the stages between receptive and productive knowledge. The learners are supposed to go through five stages in order to be able to produce the words in their speech or writing: imitation, reproduction without assimilation, comprehension, reproduction with assimilation, and production (Belyayev, 1963). The different stages the participants went through in the SW and CW conditions might have been a possible cause for their superior performance.

The results of the productive learning suggest that the SW and CW groups performed significantly better than the R and FIB in writing sentences. It might be argued that because the participants in the two aforementioned conditions made sentences in their treatment sessions, they were definitely expected to get higher marks in their learning test afterwards. However, when the productive scores are not taken into account, the receptive scores of these participants indicated a significant superiority to their counterparts.

The picture emerging from productive retention, however, was remarkably different from the learning test. Although the SW and CW groups could outperform both the R and FIB, the FIB condition did not prove to be significantly superior to the R. Contrary to the predictions of the hypothesis, this phenomenon was also observed in Hulstijn and Laufer's (2001) Dutch participants. According to what Keating (2008) suggested, it seemed that gains in productive knowledge were not able to be retained over time for tasks that were previously superior. In line with these results, in her first experiment (immediate test), Kim (2011) came to the same conclusion. Similar to the results reported for the learning condition, the participants of the CW group could not significantly outperform those of the SW. This finding corroborated that of Kim's (2011) second experiment.

Concerning the third research question, the results of the receptive and productive gains over time indicated a significant trend for the four groups in comparison with the pretest scores. This can be explained by the effectiveness of each task in vocabulary teaching and learning history. The reading tasks have been the center of attention for a long time (Gipe, 1979; Stahl, 1982). The fill-in tasks have also been declared as "highly efficient in terms of student and teacher time required" (Folse, 2006, p. 287). Furthermore, the learners will have the opportunity to study an accurate example sentence (Folse, 2006) after completing the task. The strong evaluation induced by the sentence-writing task leads to trace duration of the vocabulary knowledge (Laufer & Hulstijn, 2001) and a strong link with the output hypothesis (Swain, 1985). Similarly, the composition task has been perceived as important, due to the same properties of the SW with an extra

benefit: creation of connected pieces of discourse (Keating, 2008). However, an unsurprising phenomenon observed was the attrition of the word knowledge over time for all groups which was high enough to reach the significance level. Indeed, we had expected such a loss because the participants had no exposure to the target items in the time interval between the two post-tests (Hulstijn, 2001; Keating, 2008).

The fourth research question was an attempt to illustrate any qualitative difference among the groups over time. The percentages of the unknown, partially known, receptive, and productive knowledge were the basic criteria for such analysis. Subsequently, the superiority of the CW and SW conditions in receptive and productive gains (with a slight superiority of the CW) was indicated. Creating connected pieces of discourse might have led to more complicated processing of the target items than disjointed writing (Keating, 2008) and also "more mental effort" (Swain, 1995, p. 126) in the present context. Laufer (2003) also considers the composing task as a good option for word retention.

Considering the degree of increase in the partially known words, it was observed that the SW and CW conditions could outperform the other groups in this regard respectively. This was natural because the degree of unknown words in these conditions reached the lowest level. It seemed that the participants in these groups were able to notice (Schmidt, 1990, 1995, 2000) the target items more than the other ones. Although this noticing did not lead to establishing the form and meaning connection, at least it could create a trace of the word forms in the mind of the participants.

9. Pedagogical Implications

With regard to the implications of the present study for the teachers, they can introduce the SW and CW tasks into their classroom along with encouraging the learners to do dictionary-look up jobs. This might not only lead to more productive use of the language, but also a correct example sentence can often be found in the dictionary as a good guide (Nasrollahy Shahry, 2010). In addition, as Hulstijn and Laufer (2001) suggested, more involving tasks seem to be suitable for more important words. However, less important and easier words can be well practiced by lower involving tasks. The high-frequency words of English, such as those listed in the *General Service List of English Words* (West, 1953), the *Academic Word List* (Coxhead, 1998) and the *University Word list* (Xue & Nation, 1984) seem to be the best matches for such a rich instruction. Inasmuch as the involvement index of each task is amendable to manipulation, simple but more involving tasks can be designed containing the important words the learners might need in the course of their study. Another important implication of the

present study is concerned with the rehearsal issue which is not usually given enough time and effort it deserves. Therefore, great energy is called for reviewing the items in the succeeding sessions to ascertain the knowledge is consolidated.

Concerning the implication of the present experiment for materials writers, it seems a wise decision to include the SW and CW tasks in the course books. Although unusual at the beginning, the benefits associated with these tasks seem to pay off well. In addition, more theoretical-based tasks considering the new hypothesis can be designed with a firmer foothold than the intuition-based ones (although these tasks were successful). With regard to the rehearsal issue, allocating very small review units that engage the participants in high or even medium levels of processing might lead to such aim.

The final implication of the present study addresses the issue of assessment. Although administered thoroughly with a practice-based view thus far, the new hypothesis might give a theory-based standing to vocabulary testing purposes. Considering a very small testing section for sentence or composition writing might seem a step forward with respect to the focus on production. Indeed, the inclusion of this part might give a strong impression to the learners and this backwash effect might engage them (especially advanced learners) in the life-time process of productive learning.

References

Anderson, J. (1985). *Cognitive psychology and its implications* (2nd Ed.). New York: W. H. Freeman & Co.

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research, 20*, 303-334.

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research*, *22,* 487-497.

Belyayev, B. V. (1963). *The psychology of teaching foreign languages.* Oxford: Pergamon Press.

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

Bromberg, M., Liebb, J., & Traiger, A. (1996). *504 absolutely essential words* (4th ed.). Woodbury, NY: Barron's Educational Series, Inc.

Cambridge advanced learner's dictionary (2008). (3rd Ed.). Cambridge: Cambridge University Press.

Coxhead, A. (1998). *An academic word list.* Victoria University of Wellington: New Zealand.

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

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology, 69,* 157-169.

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40,* 273-293.

Fowler, W. S., & Coe, N. (1976). *Nelson English language tests.* (Book 2, Intermediate). London: Butler & Tanner Ltd.

Gass, S. M. (1988). Second language vocabulary acquisition. *Annual Review of Applied Linguistics, 9,* 92-106.

Gipe, J. (1979). Investigating techniques for teaching word meanings. *Reading Research Quarterly*, *14,* 624-644.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. Arnaud & H. Bejoint (Eds.). *Vocabulary and applied linguistics* (pp. 113-125). London: Macmillan.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 258-286). New York: Cambridge University Press.

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

Jing, L., & Jianbin, H. (2009). An empirical study of the involvement load hypothesis in incidental vocabulary acquisition in EFL listening. *Polyglossia, 16,* 1-11.

Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research, 11,* 149-158.

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

Kim, Y. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58,* 285-352.

Laufer, B. (1998). The development of passive and active vocabulary: Same or different? *Applied Linguistics, 19,* 255-271.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review, 59,* 567-587.

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

Longman dictionary of contemporary English (2009). (5th Ed.). Harlow, Essex: Longman.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.). *Vocabulary: Description, acquisition, and pedagogy* (pp. 109-121). Cambridge: Cambridge University Press.

Min, H. T. (2008). EFL vocabulary acquisition and retention: Reading plus vocabulary enhancement activities and narrow reading. *Language Learning, 58,* 73-115.

Nasrollahy Shahry, M. N. (2010). *The effect of receptive and productive vocabulary learning through reading and writing sentences on vocabulary acquisition*. Unpublished master's thesis, Shahid Beheshti University, Tehran, Iran.

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 151-171.

Oxford advanced learner's dictionary (2005). (7th Ed.). Oxford: Oxford University Press.

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 2,* 9-29.

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

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129-158.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.). *Attention and awareness in foreign language learning* (pp. 1-65). Hawai'i: University of Hawai'i Press.

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

Stahl, S. A. (1982). Differential word knowledge and reading comprehension. *Dissertation Abstracts International, 43,* 1517A.

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. 235-252). Rowley, MA: Newbury House.

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

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step toward language learning. *Applied Linguistics, 16,* 371-391.

West, M. (1953). *A general service list of English words*. London: Longman.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3,* 215-229.