Revisiting the Involvement Load Hypothesis: Awareness, Type of Task and Type of Item

Ana Martínez-Fernández
Georgetown University

1. Introduction

A growing number of studies in recent years (Hulstijn, 1992; Jacobs, Dufon, & Hong, 1994; Watanabe, 1997; Rott, Williams & Cameron, 2002; Rott & Williams, 2003; Rott, 2005) have examined the effectiveness of several techniques to promote incidental vocabulary learning through reading, such as glossing (i.e., providing the meaning of obscure words in the margins of a text). Because texts provide learners with a rich input where lexical items are highly contextualized, the addition of some kind of lexical intervention might further nurture lexical development. However, studies investigating the effectiveness of different lexical intervention tasks during reading have led to conflicting results (e.g., Hulstijn, 1992; Watanabe, 1997). According to the Involvement Load Hypothesis proposed by Laufer and Hulstijn (2001), incidental tasks that trigger need, search and evaluation of the meaning of unfamiliar words will lead to higher vocabulary learning than those which do not trigger such processes. The notion of 'involvement load' includes both motivational (e.g., need) and cognitive components (e.g., search and evaluation). The researchers state that incidental tasks with a higher degree of involvement load are more conducive to the type of processing that is deemed crucial for learning. This hypothesis has important pedagogical implications, since it allows us to manipulate task features and predict what tasks will be more effective. However, more empirical evidence is needed in order to support it.

In the present study, I interpret 'incidental learning' as a learning condition in which learners are processing language for meaning rather than for form (i.e., their goal is text comprehension rather than vocabulary learning) and unintentionally learn targeted forms and/or their meanings (Hulstijn, 2001; Robinson, 2002). Within this framework, learners may or may not pay attention to words and become aware of them while they are reading for meaning. Therefore, the notion of incidental learning is distinct from the notion of implicit learning, which takes place outside of awareness. While implicit learning can be incidental only, explicit learning can be both intentional and incidental. This view is different from others, where incidental learning is considered to occur when the object of learning is not the focus of attention (cf. Gass, 1999, for a discussion on different views of incidental vocabulary learning).

Most studies that are premised on the role of involvement, attention and depth of processing in incidental vocabulary learning (Watanabe, 1997; Hulstijn & Laufer, 2001) have rarely employed process measures, such as think-aloud protocols. While attempts to measure different levels of language processing and awareness have been made in cognitive psychology and other areas of SLA, this task still needs to be undertaken in studies on incidental vocabulary learning. Moreover, many of these studies assume that learning will be incidental if learners are not instructed to learn the targeted words. However, in most cases no attempt is made to ensure that learners do not expect a vocabulary test and do not intend to learn the targeted words.

Finally, while studies on intentional L2 vocabulary learning have showed that type of word – such as concrete vs. abstract nouns- might have an effect on vocabulary learning (De Groot & Keijzer, 2000), this issue remains mainly unexplored in the incidental vocabulary learning strand. The present study sought to investigate, within an attentional framework, the effect of three tasks with different degrees of involvement load on incidental L2 vocabulary development and text comprehension. In addition, it investigated different levels of awareness induced by the tasks, and addressed type of item —concrete vs. abstract nouns— as a potential variable affecting incidental vocabulary development.

2. Literature review

2.1. Levels of processing, levels of awareness and the Involvement Load Hypothesis

The concept of 'levels of processing' was proposed in the cognitive psychology field by Craik and Lockhart (1972), who suggested that remembering information depends not only on having attended to it during its occurrence or having rehearsed it after its occurrence, but also on how deeply it is processed. Laufer and Hulstijn (2001) applied this notion to the SLA field, giving rise to the Involvement Load Hypothesis in the incidental L2 vocabulary learning research strand. From a different perspective, attentional models of SLA (e.g., Robinson, 1995; Schmidt, 1990) have proposed the notion of 'levels of awareness' (Schmidt, 1990, 1995, 2001). The notion of levels of awareness has important similarities with recent views of levels of processing that relate depth of processing to elaboration and high degree of consciousness (Craik, 2002).

'Awareness' has been defined as "a particular state of mind in which an individual has undergone a specific subjective experience of some cognitive content or external stimulus" (Tomlin & Villa, 1994: p. 193). According to Schmidt (1990, and elsewhere), attention is the mechanism that controls access to awareness, and awareness can be operationally defined as ability for verbal report. He distinguishes two levels of awareness: a low level, 'noticing', and a high level, 'understanding', which involves metalinguistic awareness. The Noticing Hypothesis posits that attention is necessary for noticing, and that noticing is necessary and sufficient for intake, whereas understanding leads to deeper learning. Empirical evidence for levels of awareness has been found in several studies employing think-aloud protocols to measure awareness of morphological and syntactic structures with an on-line procedure (Leow, 1997; Rosa & O'Neill, 1999; Rosa & Leow, 2004). In a study of incidental L2 vocabulary learning, Bowles (2004) investigated whether glossing targeted words (in either computer or paper-and-pen format) promoted awareness of those words significantly more than an unglossed text. Results showed that gloss conditions reported noticing of the targeted words significantly more than the control group, but no higher levels of awareness were found.

Building on the levels of processing framework and the attentional models in SLA, Laufer and Hulstijn (2001) proposed the Involvement Load Hypothesis. In line with cognitive psychology research, Hulstijn (2001) defines depth of processing as elaboration and amount of attention: "Processing new lexical information more elaborately (e.g., by paying attention to the word's pronunciation, orthography, grammatical category, meaning and semantic relations to other words) will lead to higher retention than by processing new lexical information less elaborately (e.g., by paying attention to only one or two of these dimensions)" (p. 270). Laufer and Hulstijn (2001) proposed the notion of 'involvement' as an operationalization for the construct of depth of processing in SLA. The Involvement Load Hypothesis posits that incidental tasks that induce higher involvement are conducive to the type of processing that is deemed crucial for vocabulary retention. The notion of involvement includes three task-specific components: a motivational component, 'need', and two cognitive components, 'search' and 'evaluation'. 'Need' is defined as "the drive to comply with task requirements, whereby the task requirements can be either externally imposed (moderate need, +N) or self-imposed (strong need, ++N)" (p. 14). 'Search' (+S) and 'evaluation' (+E) require allocating attention to form-meaning relationships. 'Search' is defined as the attempt to find the meaning of an unknown word when the meaning is not provided; according to the researchers, the search process may include a variety of strategies, such as contextual guessing, consulting a dictionary, or asking the teacher. 'Evaluation' involves "a comparison of a given word with other words, a comparison of a specific meaning of a word with its other meanings, or combining the word with others in order to assess whether a word (i.e. a form-meaning pair) does or does not fit its context" (p. 14); the authors also distinguish 'moderate evaluation' (+E), when words being evaluated must fit in a given context, and 'strong evaluation' (++E), when words being evaluated must be combined with additional words in an original context, that is, a context created by the learner. Depending on the presence or absence of these components, tasks may induce a different degree of involvement on the part of the learner that will lead to noticing and elaborate processing of the words to some extent, and that will affect vocabulary retention as a result. Finally, they suggest that involvement may also be influenced by type of word (e.g., low involvement may be sufficient for learning easy words, but not for difficult ones), and that the involvement components may not be equally important for vocabulary learning.

Hulstijn and Laufer (2001) conducted two parallel experiments in two countries to test their hypothesis. In this study, three intact classes of advanced university learners of English in the Netherlands (N=87) and three intact classes with similar participants in Israel (N=99) were randomly assigned to one of three conditions varying in the involvement load induced by the task completed: (a) in the gloss condition [+N, -S, -E], participants read a text with L1 marginal glosses for ten targeted words, and answered ten multiple-choice comprehension questions; (b) in the fill-in condition [+N, -S, +E], participants read the same text and answered the same questions, but the targeted words were deleted from the text, leaving ten blanks that they had to fill in choosing the words from a list that contained 15 words with their L1 translations and L2 explanations; (c) in the writing condition [+N, -S, ++E], participants were asked to write a composition using the targeted words, for which grammatical category, L2 explanation, example, and L1 translation were provided. Vocabulary retention was measured by immediate and delayed posttests in which participants were asked to provide an L1 translation or an L2 explanation for the targeted words. Participants were informed that they would have to complete a comprehension test, and were not instructed to learn the targeted words.

The results of a 3 x 2 ANOVA indicated that the writing condition yielded significantly higher retention than the fill-in and gloss conditions in both experiments, while the fill-in group showed significantly higher retention than the gloss condition in one experiment but not in the other. The authors concluded that the hypothesis was fully and partially supported in the experiments conducted in Israel and the Netherlands respectively. However, these results can be questioned in light of the following methodological issues: (a) experimental tasks differed not only in the degree of evaluation (-E/+E/++E), but also in input vs. output orientation, and amount and quality of information provided with the targeted words in each task; (b) there was no control group; (c) process measures, such as think-aloud protocols, were not employed to ensure that tasks induced the involvement load predicted; (d) no pretest was used; instead, the likelihood of target-word familiarity was assessed in a pilot study, and prior knowledge was controlled via a post-exposure questionnaire; (e) time on task was not measured or controlled; (f) targeted items included expressions and words of different classes, so that word type was not held constant; (g) there was no randomization of participants; and (h) retention was measured by a production task only. Furthermore, participants' level of proficiency was taken as the same in both countries, but is not adequately described; additional information about their proficiency might indicate whether this factor played a role in the different results found in each experiment (i.e., significant vs. no significant difference between [+N, -S, -E] and [+N, -S, +E] conditions). Thus, further research should address these issues to find support for the Involvement Load Hypothesis.

2.2. Type of task and type of item

Most studies on traditional glosses have shown that glossing has a positive effect on text comprehension as well as on vocabulary learning, although the vocabulary gain is generally very low (e.g., Jacobs, Dufon, & Hong, 1994; Bowles, 2004). Recently, a number of studies have explained this low gain by arguing that glosses may preclude readers from inferring and guessing from context (Hulstijn, 1992; Watanabe, 1997; Hustijn & Laufer, 2001; Rott, Williams, & Cameron, 2002; Rott & Williams, 2003; Rott, 2005). According to them, inferring requires a mental effort that may result in higher vocabulary retention. For this reason, researchers have investigated different unobtrusive procedures that are assumed to require a mental effort on the learners' part, but that nevertheless guide them to get words' meanings without interrupting the reading process. One of these procedures is the multiple-choice glosses task. While traditional single glosses provide a translation, a synonym or a definition in the margin of the text, multiple-choice glosses display several translations from which the reader chooses. Findings regarding the effectiveness of multiple-choice glosses, however, are not conclusive. Some studies did not find a positive effect of multiple-choice gloss conditions when compared to control (Hulstijn, 1992) or to single gloss conditions (Hulstijn, 1992; Watanabe, 1997); Rott, Williams and Cameron (2002) found a positive effect for multiple-choice glosses on immediate posttests when compared to control and L2 reconstruction conditions (i.e., periodic L2 text reconstruction with opportunities to recheck input, and combined treatment involving both reconstruction and glosses), but not on delayed posttests; and Rott and Williams (2003) found a positive effect for a condition combining multiple-choice glosses and L2 text reconstruction when compared to an L2 text reconstruction-only condition, but did not isolate the effect of the glosses.

Recently, Rott (2005) compared the effect of single and multiple-choice glosses on processing behavior, vocabulary learning and text comprehension. In her study, 10 third-semester learners of German were randomly assigned to one of two conditions varying in the type of gloss provided. Participants were asked to think aloud while reading a text in which targeted words occurred four times, and were glossed in their first occurrence. Findings revealed a positive effect for multiple-choice glosses on 4-weeks-delayed vocabulary posttests when compared to single glosses, but not on immediate posttests. Moreover, the multiple-choice gloss group used almost twice as many strategies as the single gloss group, and used not only meta-cognitive strategies (i.e., monitoring word comprehension, verbalizing the targeted words, and referring to the glosses) but also semantic-elaborative strategies (i.e., using background knowledge and context, hypothesis-testing) to establish form-meaning connections. The author concluded that tasks involving search and evaluation processes, such as the multiple-choice gloss task, might induce deeper processing, as predicted by the Involvement Load Hypothesis, and trigger form-meaning connections that result in more robust entries in the mental lexicon. However, these results should be taken with caution due to the small sample of participants. In addition, there was no control group and time on task and outside exposure between posttests were not controlled.

Finally, none of the previous studies has focused on the potential effect of word type. Studies on intentional L2 vocabulary learning, however, suggest that word type might actually have an effect on vocabulary development (Ellis & Beaton 1993; Van Hell & Candia Mahn, 1997; De Groot & Keijzer, 2000). Specifically, findings reveal that concrete words are learned more easily than abstract words, and that forgetting is significantly higher for abstract nouns (De Groot & Keijzer, 2000).

The present study investigates the effect of three tasks with different degrees of involvement load on L2 vocabulary development through reading, and text comprehension. It also addresses levels of awareness induced by the tasks, and word concreteness as a potential variable affecting vocabulary development. In particular, this study aimed to answer the following research questions: (1) Do tasks with different degrees of involvement load induce different levels of awareness?; (2) What effect do tasks with different degrees of involvement load have on vocabulary development, and is that effect different for concrete and abstract nouns?; and (3) What effect do tasks with different degrees of involvement load have on text comprehension?

3. Methodology

3.1. Participants

Participants were 45 native English-speaking students enrolled in college-level fourth-semester Spanish language courses (Intermediate II), 25 males and 20 females, between 18 and 21 years old. The original pool consisted of 90 participants but 45 were excluded for the following reasons: (1) they had prior knowledge of one or more targeted words, (2) they expected the vocabulary tests that followed the reading, (3) they failed to follow the instructions, (4) they had outside exposure to the targeted words between the posttests, or (5) they failed to attend all sessions.

3.2. Experimental text

The input passage was an adaptation of “Shade for Sale: A Chinese Tale” (Dresser, 1994) used by Rott (2005) and translated into Spanish by the researcher. The text was 684 words long, and was modified to accommodate the targeted words. Following Rott (2005), six words (besides the targeted words) were glossed with single glosses in all conditions to further ensure comprehension.

3.3. Targeted items

The targeted items were eight Spanish words, four concrete and four abstract nouns. Each item occurred four times in the passage. The context of the first occurrence did not provide explicit clues to its meaning, while the contexts of subsequent occurrences did. According to Rott (2005), not providing

context clues for the first occurrence helps trigger evaluation in multiple-choice gloss conditions, since learners need to reconsider their initial choice as they find contextual clues.[1]

3.4. Experimental tasks

The experimental tasks differed in the degree of involvement load due to the presence or absence of 'need' (N), 'search' (S) and 'evaluation' (E). Specifically, [S] and [E] features were manipulated in three tasks so that their potential effect would be isolated. In all tasks the targeted items were bolded, glossed or deleted only during the first occurrence. In the single gloss task — [+N, -S, -E] —, participants read the text with translation single glosses of the targeted words; they did not have to search for the meaning of the words or engage in any evaluation process. In the fill-in task — [+N, -S, +E] —, the targeted words were deleted from the text, and participants had to fill in the blanks as they read, choosing from a list of 12 words accompanied by their translations. In this condition, participants did not have to search for the meaning of the words (because it was provided), but they had to compare different possible options to evaluate what word fit best in each given context. In the multiple-choice gloss task — [+N, +S, +E] —, participants read the text with translation multiple-choice glosses. Following Rott (2005), in each multiple-choice gloss three options (the correct meaning, and two additional meanings that would make sense in the present context), and a "don't know" option were given. In this condition, participants had to search for the meaning of the targeted words (because it was not provided), and compare different possible options to evaluate what meaning fit best in the given context. In the control task, the targeted words were neither bolded nor glossed nor deleted.[2]

3.5. Procedure

The experiment was conducted in three sessions. One week before the experiment, participants completed the pretest in which they were asked to (a) provide the meaning of 30 words either in English or in Spanish, (b) write "seen" if they had seen a word before, but did not know or remember its meaning, and (c) write "no" if they had not seen a word before and did not know its meaning. In the second session, participants were randomly assigned to groups. The final sample consisted of 11 participants in the control (CT) group, 10 in the single gloss (SG) group, 12 in the fill-in (FI) group, and 12 in the multiple-choice gloss (MC) group. All of them received oral instructions to think aloud in English while reading. Specifically, they were given an example of how to think aloud while solving a math problem, and were asked to say aloud anything that came to their mind. Importantly, they were not asked to explain their choices or their thoughts about the targeted words, since this type of think-aloud protocol is more likely to affect their performance (cf. Bowles & Leow, 2005). They were informed that they would have to retell the content of the passage in English, but no information was given about the vocabulary tests. To control for time on task, they were asked to indicate the time before and after completing the treatment. Immediately after, the texts were collected, and participants were asked to retell, in writing, the content of the passage in as much detail as possible.[3] This task was followed by four vocabulary posttests in the following order: (1) a word form production test, in which participants provided L2 word forms for L1 translations; (2) a sentence and word meaning production test, in which participants provided L1 translations of the targeted words as well as L2 sentences using the words; (3) a word form recognition test, in which participants chose the correct word form from three options (the correct option, a different targeted word form, and a word that was not in the text) for each L1 translation; and (4) a meaning recognition test, in which participants chose the correct meaning

[1] The items were *arce* ("maple tree"), *recámara* ("kind of room"), *aldea* ("village"), *ganado* ("livestock"), *arrebato* ("impulse"), *sosiego* ("calm"), *codicia* ("greed") and *agudeza* ("wittiness"). However, two words (*ganado* and *agudeza*) were problematic for different reasons (e.g., prior knowledge) and were excluded from the analysis, which was conducted with 3 concrete nouns and 3 abstract nouns.

[2] Single glosses were provided for six words different from the targeted words to further ensure comprehension, as in the experimental conditions. In this way, differences between control and experimental groups could be attributable to the treatment, and not to lack of comprehension of the targeted words' context.

[3] A recall protocol is not an ideal comprehension measure, since memory may play a role in writing the content of the text. However, this measure was chosen so that results could be compared with those of Rott (2005).

from three options for each targeted word; these options consisted of the same L1 translations provided in the multiple-choice gloss condition. In both recognition tests, a "don't know option" was included, and participants were encouraged not to guess. To complete the posttests, participants were given an answer sheet that was folded so that they would not turn back to a previous task. Items were presented one by one with a PowerPoint presentation, and each exposure was timed.[4] One week later, they completed the same vocabulary tests in the following order: word form production, word form recognition, meaning production, and meaning recognition.[5] Items on each test were randomized, and presented with the same procedure as on the immediate posttest. At the end of this session, participants completed a post-debriefing questionnaire, which included questions to control for outside exposure and potential expectation of a vocabulary test.

3.6. Coding

Three sets of data were coded: think-aloud protocols, written recall protocols and sentence production data. Verbal reports in the think-aloud protocols were initially classified as noticing, following Bowles (2004), who operationalized noticing as reading the glosses out loud or any comments about the targeted words (e.g., contextual guessing, commenting lack of knowledge of a word). However, based on the data collected in this study, two different types of noticing were identified and operationalized: 'noticing of one word aspect' and 'noticing of two word aspects'. 'Noticing of one word aspect' includes (a) noticing of meaning only (e.g., verbalizing the meaning of a word, reading the glosses out loud), and (b) noticing of word form only (e.g., verbalizing the word form when commenting lack of knowledge or when trying to infer its meaning unsuccessfully). The report in (1) illustrates the former type, and reports in (2) and (3) illustrate the latter:

(1) "Alberto lived in a small *village*"
(2) "I don't know what *arce* means"
(3) "*arce* probably means wall, because there is a shade…"

'Noticing of two word aspects' refers to noticing of both word form and meaning (e.g., verbalizing or writing down the word form and the meaning in the same instance). The reports in (4-7) are examples of this type of noticing:

(4) "*arce* is definitely *tree*"
(5) "oh, *arce* means *tree*, he waters the tree"
(6) "*aldea*, it could be village or shack, but he is very rich, so it's *village*"
(7) "*village* would make sense… *aldea*".

Each word for each participant was assigned one type of noticing.[6] Another researcher coded 10% of protocols and inter-rater reliability was 95%.

Prior to coding the written recall protocols, the text was analyzed to determine two sets of propositions: a set of 16 propositions expressing global ideas, where any reference to the ideas expressed in the text by the targeted words was excluded, and a set of 8 propositions containing local ideas, that is, ideas expressed by targeted words.[7] Each idea in each protocol was then coded as

---

[4] Participants were given 15 seconds for each item on the word form production test, 30 seconds to provide the meaning and write a sentence with each item on the meaning production test, and 5 seconds for each item on the multiple-choice tests. Although timing on the multiple-choice tests was pilot-tested with two students to ensure that they could answer but not think about their answer, the post-debriefing questionnaire indicated that 46% participants had time to think about their choice and even change it.

[5] This order change on the delayed posttest was done to avoid a potential memory effect. Unfortunately, responses on the meaning production test may have been affected by the preceding recognition test.

[6] The higher level of noticing was assigned to a word as long as one instance of higher awareness was instantiated at any point during task completion.

[7] The sets of propositions were determined based on Rott (2005)'s analysis of main ideas. However, a 'main idea' such as "The poor man tries to rest in the shade of the *tree*" was transformed into a 'global idea' such as "The poor man tries to rest in the shade," thus excluding any mention to the targeted word *tree*. In addition, differently from

'recalled', 'non-recalled' or as 'comprehension error' (when there was evidence that an idea had been misunderstood). Another researcher coded 10% of data and inter-rater reliability was 96%.

Sentence production data were coded according to the number of errors in grammatical category of the targeted words. Errors were counted and taken as a percentage of total instances of use. Only sentences where the targeted word was used with the correct meaning were coded.

3.7. Scoring

In the word form production test, two points were granted for a correct answer, one point for a partially correct answer, and zero otherwise.[8] On the meaning production and both recognition tests, one point was granted for a correct answer, and zero otherwise. The word form production and recognition tests consisted of 16 items (8 targeted words and 8 distracters), and the meaning production and recognition tests consisted of 8 items (8 targeted words).[9] On the text comprehension test, one point was granted for each global idea recalled, and two points for each local idea recalled, so that numbers could be directly compared.

4. Results

4.1. Preliminary analyses

Before analyzing the data to answer the research questions, several preliminary steps were taken to ensure the validity and reliability of the results. First, a one-way ANOVA was conducted on time on task data to determine whether this variable might interact with the independent variables investigated in this study. The results revealed no significant differences between the groups, $F(3, 44) = 1.097$, $p = .361$. Second, a one-way ANOVA was run on scores obtained in the multiple-choice gloss and fill-in-the blank tasks embedded in the reading. Results showed no significant differences between these two groups, $F(1, 23) = .385$, $p = .542$. Consequently, differences on the posttests cannot be attributable to differences on either time or performance on these tasks. Third, the reliability coefficients for the posttest tasks were computed using Cronbach's alpha. For the word form production, meaning production, and word form recognition tasks, reliability was medium-low (.62, .67, .62, respectively), and low for the meaning recognition task (.49).

Finally, the think-aloud protocols were analyzed to ensure that participants were representative of their condition. Specifically, this analysis focused on two aspects: (a) whether participants showed evidence of intentional learning; and (b) whether participants engaged in search and evaluation processes according to the involvement load of the task they performed. No evidence for intentional learning or expectation of subsequent vocabulary tests was found.[10] With regard to the processes induced by the task conditions, it was especially relevant to determine whether the fill-in task triggered similar evaluation processes to the multiple-choice gloss task. It could be argued that the fill-in task induced stronger evaluation because participants had to evaluate 12 options for each blank, while participants in the multiple-choice gloss condition had to evaluate only 3 options for each multiple-choice question. Moreover, participants in the fill-in condition might have evaluated the meaning of some of the words several times, since for each blank all of the targeted words were possible options. However, the think-aloud protocols indicated that participants in the multiple-choice gloss condition did evaluate the words and meanings many times; since they encountered each targeted word four times in the text, they usually went back to change or confirm their initial choice based on the new contextual clues. This evaluation process, however, did not always lead them to choose the right option. In

---

Rott (2005), I decided not to include a set of 'supporting ideas', because these ideas involved a lot of specific details, and memory might easily affect their recall. Instead, a set of 'local ideas' seemed more appropriate.

[8] In the word form production test, a maximum of two orthographic mistakes in words containing more than five letters, and one mistake in words containing five or fewer letters was allowed in order to consider an answer as partially correct. In the meaning production test, no examples of partially correct answers were found.

[9] Distracters were included in the word form production and recognition tests so that the correct meaning of the targeted words would not be prompted in the meaning production and recognition tests.

[10] However, 16 participants indicated in the post-debriefing questionnaire that they expected the vocabulary test. Since this expectation might have affected their performance, they were excluded from the final sample.

contrast, participants in the fill-in condition did evaluate several options to fill in each blank, but did not go back to change their choices. Thus, although the number of options to evaluate was not held constant across tasks, both tasks were successful in triggering similar evaluation processes. However, because participants in the fill-in task were provided with the meaning of the targeted words, the involvement load of this task was lower than that of the multiple-choice gloss task.[11]

4.2. Levels of awareness

To answer the first research question, numbers of instances of each type of verbal report for each condition were submitted to a MANOVA,[12] using a one between-subjects design with three dependent variables: 'noticing of one word aspect' (-N), 'noticing of two word aspects' (+N), and 'no verbal report' (NVR). The means displayed in Table 1 show that the fill-in group, followed by the multiple-choice gloss group, reported the highest amount of noticing of two word aspects.

Table 1. Means for concurrent verbal report by Group

Verbal report	*Group*	*N*	*M*	*SD*	*SE*
[NVR]					
	[CT]	11	1.82	1.471	.504
	[SG]	10	.60	.966	.529
	[FI]	12	1.33	1.775	.483
	[MC]	12	1.00	2.132	.483
[-N]					
	[CT]	11	3.91	1.514	.426
	[SG]	10	4.50	1.581	.447
	[FI]	12	.83	1.115	.408
	[MC]	12	3.08	1.443	.408
[+N]					
	[CT]	11	.27	.467	.376
	[SG]	10	.90	1.524	.394
	[FI]	12	3.83	1.467	.360
	[MC]	12	1.92	1.240	.360

This analysis showed a significant effect for Group, $\Lambda = .319$, F (6, 80) = 10.275, $p = .000$, and a large effect size ($\eta^2 = .435$). A post-hoc Scheffé test indicated the following results: (a) there were no significant differences between groups in the amount of no verbal report; (b) the amount of reported noticing of one word aspect was significantly lower in the fill-in group than in all other groups; and (c) the amount of reported noticing of two word aspects was significantly higher in the fill-in group than in all other groups, and significantly higher in the multiple-choice gloss group than in the control condition; there was no significant difference between the multiple-choice gloss and single gloss conditions. To further investigate what subtype of noticing was found in each condition, another MANOVA was conducted with four dependent variables: 'noticing of word form only', 'noticing of meaning only', 'noticing of word form and meaning', and 'no verbal report'. The means displayed in

[11] In addition, responses to the post-debriefing questionnaire suggest that participants in the multiple-choice gloss group were highly involved in the task; they often reported to feel frustrated and confused due to the uncertainty of their choices. In contrast, participants in the fill-in condition did not report such an experience, even though their performance was not more accurate than that of the multiple-choice gloss group.

[12] A MANOVA was used to reduce the likelihood to make a Type I error.

Table 2 show that the single gloss group reported the highest amount of noticing of meaning only, while the multiple-choice gloss and control conditions reported a higher amount of noticing of word form only.

Table 2. Means for type of noticing reported by Group

Verbal report	*Type of task*	*N*	*M*	*SD*	*SE*
[-N: Form]					
	[CT]	11	3.91	1.514	.366
	[SG]	10	.10	.316	.384
	[FI]	12	.83	1.115	.350
	[MC]	12	3.08	1.443	.350
[-N: Meaning]					
	[CT]	11	.00	.000	.213
	[SG]	10	4.40	1.506	.223
	[FI]	12	.00	.000	.204
	[MC]	12	.00	.000	.204
[+N]					
	[CT]	11	.27	.467	.376
	[SG]	10	.90	1.524	.394
	[FI]	12	3.83	1.467	.360
	[MC]	12	1.92	1.240	.360

The results yielded a significant effect for Group, $\Lambda = .031$, $F\ (9, 95) = 22.492$, $p = .000$, and a large effect size ($\eta^2 = .685$).[13] This analysis showed that the amount of reported noticing of word form only was significantly higher in the control and multiple-choice gloss groups than in the single gloss and fill-in groups, while the amount of reported noticing of meaning only was significantly higher in the single gloss group than in all other conditions.[14]

In addition, the think-aloud protocols indicated that there were differences between conditions, as well as individual differences, regarding the strategies adopted in the second, third and fourth encounters with the targeted words. In the single gloss group, participants noticed the glosses, and many transferred the meaning of the targeted words in subsequent encounters, or referred back to the glosses; others, however, skipped the words or indicated a lack of knowledge of them. In the fill-in group, participants usually transferred the meaning to subsequent encounters, and did not go back to change their previous choice. In the multiple-choice gloss group, participants evaluated the new contexts and went back to change or confirm their choices; some transferred the meaning to subsequent encounters, but others skipped the targeted words when translating the text. In the control group, participants often indicated a lack of knowledge of the targeted words, tried to guess their meaning from the context, and in fewer cases, they ignored the targeted words.[15] Generally, high-level processes such as hypothesis testing, evaluation of the context and inferences were found in the fill-in group, the multiple-choice gloss group, and to a lesser extent, in the control group.

[13] In the scale for Eta squared, .01 represents a small effect, .06 represents a medium effect, and .14 represents a large effect (differently from Cohen's *d*, where .1 is small, .6 is medium, and .8 is large).

[14] Differences in amount of noticing cannot be attributed to any difference in overall amount of thinking aloud. Typically, participants in all conditions read aloud and translated the text, or directly translated the text from the beginning to the end. Therefore, the overall amount of thinking aloud (measured in time) coincides with time on task, which was not significantly different between groups.

[15] A quantitative analysis of the strategies adopted in each encounter with the targeted words, as the one conducted by Rott (2005), might shed some light on the results found. However, this analysis is beyond the scope of this study.

4.3. Vocabulary development

To answer the second research question, separate 4 x 2 x 3 repeated measures ANOVA were conducted, using a one between-subjects and two within-subject (Word type and Time) design.

4.3.1. Word form production

The means of the scores on the word form production tests show that the fill-in group outperformed all other groups on both the immediate and delayed posttests and on both types of nouns. Table 3 shows means, standard deviations and standard errors for word form production.

Table 3: Means for word form production by Time and Word type

Time	*Group*	*Word type*						
			Concrete			*Abstract*		
		N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Immediate								
	[CT]	11	1.27	1.272	.465	.36	.809	.459
	[SG]	10	1.70	1.829	.488	1.00	1.333	.481
	[FI]	12	2.67	1.723	.445	2.17	2.406	.439
	[MC]	12	1.67	1.303	.445	.67	.888	.439
Delayed								
	[CT]	11	.55	.820	.466	.18	.603	.380
	[SG]	10	1.60	1.776	.489	.40	.843	.399
	[FI]	12	2.25	1.960	.446	1.50	2.153	.364
	[MC]	12	1.67	1.371	.446	.25	.622	.364

(Maximum for each word type: 6)

Results for word from production yielded a significant main effect for Time, F (2, 41) = 40.519, p = .000, a significant main effect for Word type, F (1, 41) = 15.401, p = .000, a significant main effect for Group, F (3, 41) = 64.110, p = .012, a significant interaction Time x Group, F (6, 41) = 3.429, p = .001, a significant interaction Time x Word type, F (2, 82) = 6.980, p = .002, no significant interaction Word type x Group, F (3, 41) = .432, p = .732, and no significant interaction Time x Word type x Group, F (6, 82) = .582, p = .582. For all significant effects and interactions a large effect size was found.[16] Results revealed that concrete nouns were produced significantly more than abstract nouns. A post-hoc Scheffé test was conducted to identify differences between groups. This test showed that only the fill-in group improved from the pretest to the immediate posttest significantly more than the control group. In addition, the loss from the immediate to the delayed posttest was not significantly different for groups, and was significantly higher for abstract nouns than for concrete nouns.

[16] A small effect size and low power are found for the interactions Word type x Group and Time x Word type x Group, indicating that if these interactions had a significant effect, it would be very small, and thus a greater sample size would be needed in order to test it. This result is found in all measures, except for word form recognition test, in which the interaction Time x Word type x Group is significant, and the effect size is large.

4.3.2. Meaning production

The means of the scores on the meaning production tests indicate that the fill-in group, followed by the single gloss group, performed higher than the other groups on both the immediate and delayed posttests and on both types of nouns. Table 4 displays descriptive statistics for meaning production.

Table 4: Means for meaning production by Time and Word type

Time	*Group*	*Word type*						
			Concrete			*Abstract*		
		N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Immediate								
	[CT]	11	.91	1.044	.909	.27	.467	.273
	[SG]	10	1.80	1317	.329	1.30	1.059	.308
	[FI]	12	2.08	1.084	.300	1.75	2.288	.281
	[MC]	12	1.58	1.669	.300	.75	.866	.281
Delayed								
	[CT]	11	1.27	1.009	.308	.36	.505	.319
	[SG]	10	2.10	1.197	.323	.90	1.197	.335
	[FI]	12	2.50	.905	.295	1.92	1.311	.306
	[MC]	12	1.33	.985	.295	.83	1.030	.306

(Maximum for each word type: 3)

Results yielded a significant main effect for Time, F (2, 41) = 140.642, p = .000, a significant main effect for Word type, F (1, 41) = 14.772, p = .000, a significant main effect for Group, F (3, 41) = 7.410, p = .000, a significant interaction Time x Group, F (6, 41) = 5.301, p = .000, a significant interaction Time x Word type, F (2, 82) = 9.479, p = .000, no significant interaction Word type x Group, F (3, 41) = .227, p = .877, and no significant interaction Time x Word type x Group, F (6, 82) = .760, p = .604. For all significant effects and interactions a large effect size was found. Results indicate that the meaning of concrete nouns was produced significantly more than the meaning of abstract nouns. A post-hoc Scheffé test showed that only the fill-in group improved significantly more than the control group. In addition, the fill-in group performed significantly higher than the multiple-choice gloss group, whose performance was not significantly different from that of the control group. The single gloss group was not significantly different from any group. Interestingly, the change from the immediate to the delayed posttest was significantly different for conditions. While the fill-in, multiple-choice gloss and control groups improved on the delayed posttest, the single gloss group did not. Overall, the gain from the immediate to the delayed posttest was significantly higher for concrete nouns than for abstract nouns.

4.3.3. Word form recognition

The means of the scores on the word form recognition tests indicate the same trend found in previous analyses: the fill-in group performed higher than the other groups on both the immediate and delayed posttests and on both types of nouns. Table 5 shows descriptive statistics for word form recognition.

Table 5: Means for word form recognition by Time and Word type

Time	*Group*	*Word type*						
			Concrete			*Abstract*		
		N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Immediate								
	[CT]	11	1.45	.934	.249	.82	.603	.281
	[SG]	10	2.20	1.033	.261	1.90	.994	.295
	[FI]	12	2.50	.674	.239	2.42	.793	.260
	[MC]	12	2.33	.651	.239	1.25	1.215	.260
Delayed								
	[CT]	11	1.18	1.079	.297	.55	.934	.310
	[SG]	10	1.90	1.197	.312	1.20	1.229	.325
	[FI]	12	2.67	.492	.285	2.00	.853	.296
	[MC]	12	1.67	1.073	.285	1.42	1.084	.296

(Maximum for each word type: 3)

Results yielded a significant main effect for Time, F (2, 82) = 154.452, p = .000, a significant main effect for Word type, F (1, 41) = 13.675, p = .001, a significant main effect for Group, F (3, 41) = 8.189, p = .000, a significant interaction Time x Group, F (6, 82) = 4.543, p = .001, a significant interaction Time x Word type, F (2, 82) = 8.085, p = .001, a significant interaction Time x Word type x Group, F (6, 82) = 2.254, p = .046, and no significant interaction Word type x Group, F (3, 41) = .219, p = .883. For all significant effects and interactions a large effect size was found. Results indicated that concrete nouns were recognized significantly more than abstract nouns. A post-hoc Scheffé test showed that the gain from the pretest to the immediate posttest was significantly higher for the fill-in group than for the control group, while the other experimental conditions were not significantly different from any condition. The loss from the immediate to the delayed posttest was significantly higher for the single gloss and control groups than for the fill-in and multiple-choice gloss groups. Finally, the loss from the immediate to the delayed posttest was significantly higher for abstract nouns than for concrete nouns in all groups, except for the multiple-choice gloss group, whose retention of abstract nouns was significantly higher than that of the other groups.

4.3.4. Meaning recognition

The means of the scores on the word form recognition tests indicate that the fill-in group, followed by the single gloss group, performed higher than the other groups on both posttests and on both types of nouns. Table 6 displays descriptive statistics for meaning recognition.

Table 6: Means for meaning recognition by Time and Word type

Time	*Group*	*Word type*						
			Concrete			*Abstract*		
		N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Immediate								
	[CT]	11	1.55	1.036	.909	1.00	.775	.273
	[SG]	10	2.30	1.160	.329	2.00	.943	.308
	[FI]	12	2.67	.651	.300	2.25	.866	.281
	[MC]	12	1.92	.996	.300	1.50	.798	.281
Delayed								
	[CT]	11	1.27	.905	.308	.91	.831	.319
	[SG]	10	2.20	1.033	.323	1.30	1.160	.335
	[FI]	12	2.42	.996	.295	2.17	1.193	.306
	[MC]	12	1.58	1.084	.295	1.33	.778	.306

(Maximum for each word type: 3)

Results yielded a significant main effect for Time, F (2, 82) = 228.906, p = .000, a significant main effect for Word type, F (1, 41) = 5.861, p = .020, a significant main effect for Group, F (3, 41) = 355.464, p = .000, a significant interaction Time x Group, F (6, 82) = 4.923, p = .000, a significant interaction Time x Word type, F (2, 82) = 3.273, p = .043, no significant interaction Word type x Group, F (3, 41) = .122, p = .947, and no significant interaction Time x Word type x Group, F (6, 82) = .524, p = .788. For all significant effects and interactions a large effect size was found, except for the interaction Time x Word type, where the effect size was medium (η^2 = .07). Results indicated that concrete nouns were recognized significantly more than abstract nouns. A post-hoc Scheffé test revealed that the gain from the pretest to the immediate posttest was significantly higher for the fill-in group than for the control group, as can be observed in Figures 1 and 2.[17]

Figure 1. Meaning recognition of concrete nouns

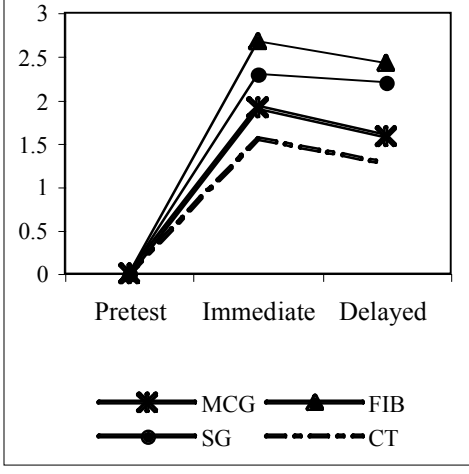
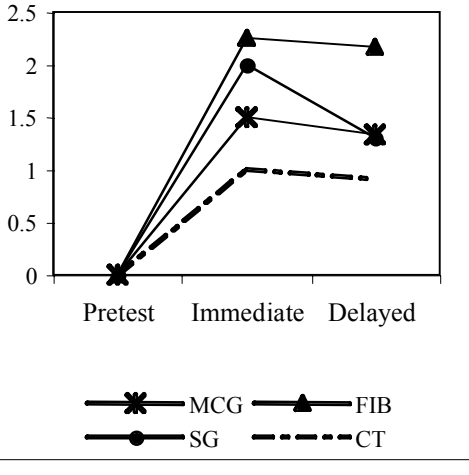


Figure 2. Meaning recognition of abstract words



[17] Due to space limitations, not all figures are included. However, Figures 1 and 2 illustrate the general pattern.

In addition, the fill-in group improved significantly more than the multiple-choice gloss group, whose performance was not significantly different from that of the control group. The single gloss group was not significantly different from any group. All groups recognized fewer words on the delayed posttest than on the immediate posttest, but this loss was significantly higher for the single gloss group than for the other groups, consistent with the pattern observed on the meaning production test. Finally, the loss from the immediate to the delayed posttest was significantly higher for abstract nouns than for concrete nouns, as was observed in all other measures.

4.3.5. Sentence production

The analysis of sentences revealed that abstract nouns, regardless of the group, were used as adjectives in 41.8% and 35% of cases on the immediate and the delayed posttest respectively, while concrete nouns were always used as nouns. The examples below illustrate this pattern:

(a) "Alberto el Rico estaba muy codicio" ("Rich Albert was very greed")
(b) "Estoy sosiega cuando tengo vacaciones" ("I am calmness when I am on holidays")
(c) "Cuando me enfado soy muy arrebato" ("When I get angry I am very impulse")

4.4. Text comprehension

To answer the third research question, the raw scores taken from the recall protocols were submitted to a separate 4 x 2 repeated measures ANOVA, using a one between-subjects and one within-subject design. The means displayed in Table 7 show that, on the one hand, the single gloss, control and fill-in groups recalled more global ideas than the multiple-choice gloss group, and that the single gloss group, followed by the fill-in group, recalled more local ideas than the multiple-choice gloss and control groups. On the other hand, the multiple-choice gloss group, followed by the control group, made more comprehension errors than the other groups.

Table 7. Means for ideas recalled and comprehension errors by Group and Idea type

Idea type	*Group*	*Measure*						
			Ideas recalled			*Comprehension errors*		
		N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Global								
	[CT]	11	12.73	1.794	.893	.64	1.120	.361
	[SG]	10	13.10	3.247	.937	.40	.843	.379
	[FI]	12	12.50	2.468	.855	.08	.289	.346
	[MC]	12	10.83	3.881	.855	1.08	1.881	.346
Local								
	[CT]	11	2.36	2.335	1.084	2.00	1.789	.461
	[SG]	10	9.60	3.864	.1.137	.00	.000	.483
	[FI]	12	7.50	4.442	1.038	1.00	1.044	.441
	[MC]	12	5.83	3.353	1.038	1.33	2.167	.441

(Maximum for each idea type: 16)

Results for number of ideas recalled yielded a significant main effect for Group, F (3, 41) = 3.718, p = .019, a significant main effect for Idea type, F (1, 41) = 158.085, p = .000, and a significant interaction Idea type x Group F (3, 41) = 9.792, p = .000. The effect size was large in all cases (η^2 =

.214, .794, and .417, respectively). The results showed that (a) all groups recalled significantly more global ideas than local ideas, (b) there was no significant difference between groups in the number of global ideas recalled, and (c) there was a significant difference between groups in the number of local ideas recalled, as Figure 3 below clearly shows. A post-hoc Scheffé test revealed that the single gloss group recalled significantly more local ideas than the multiple-choice gloss and the control groups, and that there was no significant difference between the fill-in and the single gloss groups. Results for comprehension errors yielded a significant main effect for Group, F (3, 41) = 4.979, p = .005, a significant main effect for Idea type, F (1, 41) = 6.769, p = .014, and no significant interaction Idea type x Group F (3, 41) = 1.745, p = .173. The effect size was large for the Group and Idea type effects (η^2 = .267, 142, respectively) and medium-large for the interaction effect (η^2 = .113). The results indicated that all groups made more comprehension errors in local ideas than global ideas. A post-hoc Scheffé test indicated that the single gloss group made significantly fewer errors than the multiple-choice gloss and control groups, and that the fill-in group made significantly fewer errors than the multiple-choice gloss group. There were no significant differences between the single gloss and fill-in groups, or between the multiple-choice gloss and control groups. Figure 4 illustrates these results.

Figure 3. Recall of ideas

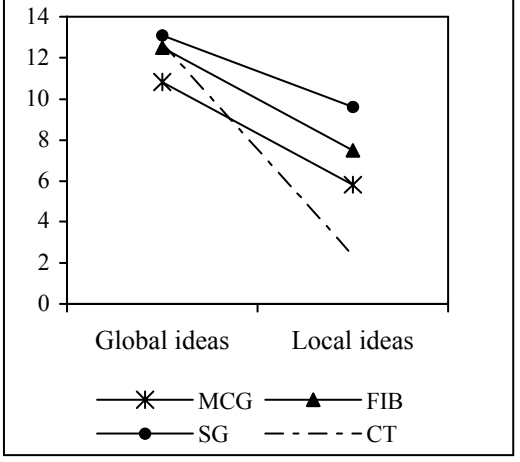
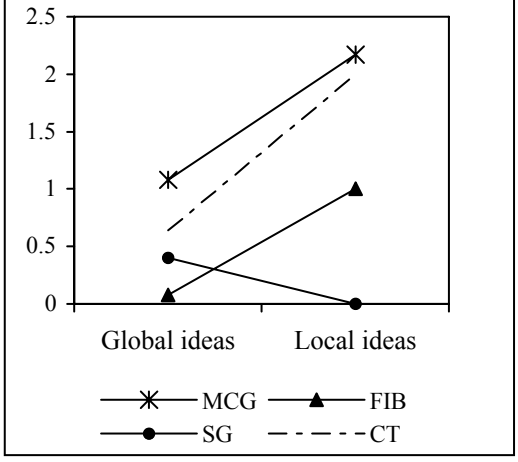


Figure 4. Comprehension errors



5. Discussion, limitations and further research

The first research question investigated whether tasks with different degrees of involvement load induced different levels of awareness as measured by think-aloud protocols. In this study I identified three types of 'noticing', operationalized as ability for verbal report: noticing of word form only, noticing of meaning only, and noticing of both word form and meaning. Following Bowles (2004), the first two types of noticing —word form only and meaning only— constitute one level of awareness. However, there are a number of issues that should be taken into account when assessing noticing of one word aspect. First, noticing of word meaning only does not necessarily imply that the word form is not noticed. Learners in single gloss conditions, for example, probably pay attention to the word form before realizing their lack of knowledge and/or looking up the gloss. However, because learners are reading for meaning and do not have to complete any lexical task, it is likely that some of them will process the word form and its meaning differently. Noticing of meaning only, as a concept, does not exclude attention to the word form, but implies deeper processing of the meaning of the word. Second, an important limitation of the measure of awareness used in this study is that learners may not verbalize all their thoughts, and thus, a lack of verbalization cannot be interpreted as a lack of awareness; consequently, the operationalization of noticing of meaning only does not exclude the possibility that both word form and meaning are deeply processed or noticed. This limitation also holds for the operationalization of noticing of word form only, although in the data verbalizations of word form only were usually accompanied by explicit verbalizations of lack of knowledge of the meaning of the word.

Third, it is important to distinguish awareness from high-level processes such as hypothesis testing or evaluation. Although they may be related in some cases, awareness is not isomorphic with such processes. For example, noticing of word form only was operationalized as verbalizations of the word form, such as those produced when commenting lack of knowledge or when making wrong inferences about the meaning of the word. Clearly, the process of a learner who indicates lack of knowledge of a word and moves on is distinct from that of a learner who evaluates the context and tries to infer its meaning. These processes can be referred to as low and high-level processes, respectively. However, in both cases there is awareness of a word form and a lack of awareness of the meaning of that word. Thus, one level of awareness can be related to two different processes.

Noticing of both word form and meaning was interpreted in this study as a high level of awareness. Schmidt (1995) makes a distinction between two levels of awareness, a low level ―noticing― which involves a cognitive registration of a form, and a higher level ―understanding―, which involves awareness of more word aspects, such as understanding the meaning and syntactic features of the word. Within this framework, it can be interpreted that noticing of only one word aspect, either the word form or the meaning, is a cognitive registration of a form that constitutes a low level of awareness. In turn, noticing of both word form and meaning can be interpreted as a higher level of awareness, since it involves a cognitive registration of more word aspects.[18] Therefore, this study supports previous findings of studies on levels of awareness employing morphological and syntactic structures (Leow, 1997; Rosa & O'Neill, 1999; Rosa & Leow, 2004). As in these studies, the higher level of awareness was found in all conditions, and was usually related to high-level processes such as hypothesis testing, inferring, and context evaluation (except for the single gloss condition, where no evidence of such processes was found). These processes, however, not always led to high awareness, as has been mentioned above.

Findings in this study show that tasks with different degrees of involvement load induced different levels of awareness, although not in the direction predicted by the Involvement Load Hypothesis. The fill-in task induced significantly higher awareness than all other conditions, including the multiple-choice gloss task, which had the highest degree of involvement load; the multiple-choice gloss task induced significantly higher awareness than the single gloss and control tasks. These results suggest that (a) the fill-in and multiple-choice gloss conditions led to qualitatively different processing, which seems to be unrelated to the degree of evaluation (both conditions triggered evaluation processes, as revealed by the think-aloud protocols), and (b) the 'evaluation' component may be more effective than the 'search' component in inducing high awareness, at least in incidental vocabulary learning tasks where the main goal is understanding a text.

The second research question investigated whether tasks with different degrees of involvement load had a different effect on vocabulary development, and whether that effect was different for concrete and abstract nouns. Results indicate that the tasks used in this study did have a different effect on vocabulary gain, but do not support predictions made by the Involvement Load Hypothesis, as illustrated by the following significant effects found: (a) the fill-in group outperformed the control group on the word form production and recognition tests, and outperformed both the control and the multiple-choice gloss groups on the meaning production and recognition tests; (b) the multiple-choice gloss group was not different from either the control group or the single gloss group on any test; and (c) the single gloss group was not different from any other group on any test. Therefore, higher vocabulary development occurred in the fill-in group than in the multiple-choice gloss group. Since the fill-in group also reported higher awareness than any other group, it seems plausible that high awareness might have led to greater learning,[19] as it was found in previous studies employing grammatical structures (Leow, 1997; Rosa & O'Neill, 1999; Rosa & Leow, 2004). The fact that the fill-in and single gloss groups were not significantly different on any test does not necessarily contradict this explanation. The single gloss group did not report a significant amount of high awareness, but did report a high amount of noticing the meaning of the targeted words during the first encounter. Since the targeted words appeared three more times in the text, it is likely that some participants noticed the word forms in

[18] This higher level of awareness might not be interpreted as 'understanding', since this notion involves also awareness of syntactic features. In addition, it should not be confused with the notion of 'form-meaning connection', which does not necessarily involve awareness (VanPatten, Williams, Rott, & Overstreet, 2004).

[19] These results, however, should be taken with caution, since the correlation between levels of awareness and vocabulary development was not addressed in this study.

subsequent encounters (in addition to the meaning, which was transferred sometimes from the first encounter) without reporting such noticing.[20] In that case, the single gloss condition might have induced a higher level of awareness than what it is possible to report (although not as high as to significantly outperform the multiple-choice gloss group).[21] Interestingly, no significant differences between the multiple-choice gloss and control groups were found, supporting results reported by Hulstijn (1992) but contradicting findings reported by Watanabe (1997). The multiple-choice gloss task in Hulstijn (1992) and the present study included either four choices or three choices and a "don't know" option, while in Watanabe (1997) it only included two choices. It is possible that the more alternatives participants are given in this task, the more confused they become, and thus the more difficult for them to notice what the correct meaning is.

Regarding the effect of time on vocabulary development, findings revealed a significantly higher loss (or lower gain)[22] from the immediate to the delayed posttests in the single gloss group than in all other groups on both the meaning production and recognition tests, as well as a significantly higher retention of abstract words in the multiple-choice gloss group on the word form recognition test. These results support findings reported by Rott (2005), where the multiple-choice gloss group showed higher retention on a four-weeks-delayed posttest than the single gloss group. As has been pointed out above, the single gloss condition was the only one in which high-level processes were not found when analyzing the think-aloud protocols. While these processes might not be necessary to become aware of the targeted items, it is possible that they have an important effect on retention. This might explain the significantly lower retention experienced by the single gloss group, and, to a lesser extent, by the control group, in which high-level processes were found but not as frequently as in the multiple-choice gloss and fill-in conditions.

With respect to the effect of type of item, findings show that all groups performed significantly higher on concrete than abstract nouns, and that retention was significantly lower for abstract than for concrete nouns on all tests, as previous studies have found (e.g., De Groot & Keijzer, 2000). These results should be taken with caution due to the low number of items and the lack of control for a number of factors, such as degree of difficulty of pronunciation, number of syllables, relevance, inferability, and syntactic function of the targeted words. Results also revealed that abstract nouns were often used as adjectives in sentences that participants produced.[23] However, when providing their meaning, participants always provided English nouns instead of adjectives. Therefore, the difficulty with abstract nouns seems to be related to the ability to use the word in a sentence, and not necessarily to knowledge of the word category. In any case, this result suggests that some syntactic features of the words might require a higher level of awareness for them to be internalized.

The third research question investigated what effect tasks with different degrees of involvement load had on text comprehension as measured by number of global and local ideas recalled and number of comprehension errors in an L1 recall task. Findings showed that all groups recalled a high number of global ideas. Because the number of global ideas not recalled was very low, it can be concluded that all groups had a high global comprehension of the text. In contrast, the results for local ideas reflect to some extent the trends found in vocabulary development: there were no significant differences between the multiple-choice gloss and control groups, or between the single gloss and fill-in groups, and the single gloss group significantly outperformed both the multiple-choice gloss and control groups.

---

[20] An anonymous reviewer suggests that noticing of meaning only may be interpreted as high awareness, involving reported noticing of meaning and non-reported noticing of word form. However, I believe that word frequency played a role in noticing the word form in this study. This frequency effect would also explain the high amount of noticing reported by the control group. Future studies will have to investigate conditions involving non-frequent targeted words, and determine whether any potential effect may or may not be related to differences between noticing of meaning only and noticing of both word form and meaning, as operationalized in this study.

[21] The lack of significant difference between single gloss and multiple-choice gloss conditions on the immediate posttests supports previous findings reported by Watanabe (1997) and Rott (2005).

[22] The gain on both the meaning production and recognition tests may be due to a carry-over effect.

[23] The word most often used as an adjective was *sosiego*. Because the L1 translation for this word (*calm*) can be both a noun and an adjective, it could be argued that incorrect category use was induced by the translation provided. However, this explanation would not account for the other targeted words (including the word excluded, which was also used as an adjective).

However, the number of local ideas recalled was significantly lower than that of global ideas. Therefore, it cannot be claimed that the multiple-choice gloss and control groups had lower local comprehension based on the fact that they recalled a lower number of ideas. For this reason, results for comprehension errors may provide a more complete picture. Findings show that there was no difference between the number of global and local comprehension errors. Interestingly, the single gloss group made significantly fewer errors than the multiple-choice gloss and control groups, and the fill-in group made significantly fewer errors than the multiple-choice gloss group. There was no significant difference between the single gloss and fill-in conditions, or between the multiple-choice gloss and control conditions. Overall, the number of errors was not high; however, since the number of local ideas recalled was low, it is not possible to know whether the local ideas not recalled were misunderstood. In sum, these results show that the multiple-choice gloss task may have a negative effect on global and local comprehension when compared to single gloss and fill-in conditions, although it does not have a negative effect when compared to the control condition. In contrast, the single gloss group, followed by the fill-in group, recalled more global and local ideas, and made fewer comprehension errors than the multiple-choice gloss condition.

To conclude, I did not find support for the Involvement Load Hypothesis in this study: tasks with higher degree of involvement load did not lead to deeper processing, defined as high awareness, and did not lead to higher vocabulary development. A number of differences between this study and the studies conducted by Hulstijn and Laufer (2001), such as the use of think-aloud protocols or the proficiency level of participants (intermediate vs. advanced), may be claimed to play a role in the different results found. However, since the highest involvement load was confounded with output-orientation, it seems likely that output, and not evaluation, had the highest effect on vocabulary development in those studies. When input-orientation tasks are compared, search and evaluation do not seem to have such a positive effect. Finally, the present study shows that type of word ―concrete vs. abstract nouns― might play a role in vocabulary development, and that tasks may impact text comprehension in a different way. A pedagogical implication of the findings reported in this study is that fill-in tasks for words that appear frequently in a text may be more beneficial than multiple-choice glosses tasks.

Further research should use more refined vocabulary and comprehension measures, control variables affecting type of item, conduct an in-depth analysis of strategies used by learners, and address potential reactivity of thinking aloud. Although think-aloud protocols have proved to be a successful tool to operationalize awareness, their potential reactivity might affect learners' performance in some cases. In SLA research, one study employing L2 think-aloud protocols has found negative reactivity (Sachs & Polio, 2007). However, most studies employing L1 think-aloud protocols have found reactivity on time on task but not on performance (Leow & Morgan-Short, 2004; Bowles & Leow, 2005; Sachs & Suh, 2007). To date, only one study suggests that thinking aloud in the first language may have a positive effect in less explicit conditions and no effect in more explicit conditions, which indicates that reactivity might also depend on type of task (Sanz, Lin, Lado, Bowden & Stafford, in press). None of these studies, however, has addressed the issue of reactivity on vocabulary development. Finally, a number of issues, such as levels of awareness, relationship between levels of awareness and vocabulary development, effects of high-level processes on retention, and effects of both quality and quantity of exposure should be further investigated.

Acknowledgements

I would like to thank Ronald P. Leow, Grant Armstrong and the anonymous reviewers for their valuable comments and insights. I also thank Cristina Sanz for her helpful advice, and Susan Rott for letting me use her materials.

References

Bowles, M. A. (2004). L2 glossing: To CALL or not CALL. *Hispania, 87*(3), 541-552.

Bowles, M. A. & Leow, R. P. (2005). Reactivity and type of verbal report in SLA research methodology. *Studies in Second Language Acquisition, 27*(3), 415-440.

Craik, F. I. M. (2002). Levels of processing: Past, present… and future? *Memory, 10* (5/6), 305-318.

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.

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting, *Language Learning, 50*(1), 1-56.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*, 559-617.

Gass, S. (1999). Incidental vocabulary learning. *Studies in second language acquisition, 21*, 319-333.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. 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, *Cognition and Second Language Instruction*, (pp. 258-286). Cambridge: Cambridge University Press.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539-558.

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading, 17*(1), 19-28.

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

Leow, R. P. (1997). Attention, awareness and foreign language behavior. *Language Learning, 47*, 467-505.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*(1), 35-57.

Robinson, P. (1995). Attention, memory and the "noticing" hypothesis. *Language Learning, 45*(2), 283-331.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA. In P. Robinson, *Individual Differences and Instructed Language Learning* (pp. 211-266) Philadelphia/Amsterdam: John Benjamins.

Rosa, E., & Leow, R. P. (2004). Awareness, different language conditions, and second language development. *Applied Psycholinguistics*, 25, 269-292.

Rosa, E., & O'Neill, M. D. (1999). Explicitness, intake, and the issue of awareness. Another piece to the puzzle. *Studies in second language acquisition, 21*(4), 511-556.

Rott, S. (2005). Processing glosses: A qualitative exploration of how form-meaning connections are established and strengthened. *Reading in a Foreign Language, 17*(2), 95-124.

Rott, S., & Williams, J. (2003). Making form-meaning connections while reading: A qualitative analysis of word processing. *Reading in a Foreign Language*, 15(1), 45-75.

Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research, 6*, 183-222.

Sachs, R. & Polio, C. (2007). Learners' uses of two types of written feedback on a L2 writing revision task. *Studies in Second Language Acquisition*, *29*(1), 67-100.

Sachs, R., & Suh, B. (2007). Textually enhanced recasts, learner awareness, and L2 outcomes in synchronous computer-mediated interaction In A. Mackey (Ed). *Conversational Interaction in Second Language Acquisition.* Oxford Applied Linguistics.

Sanz, C., Lin, H., Lado, B., Bowden, H. W., & Stafford, C. A. (in press). Concurrent verbalizations, pedagogical conditions, and reactivity: two CALL studies. *Language learning, 59*(1).

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

Schmidt, R. (1995). Consciousness and foreign language learning. In R. Schmidt (Ed.), *Attention and Awareness in Foreign Language Learning* (pp. 1-63). University of Hawai'i at Manoa: Second Language Teaching and Curriculum Center.

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

Tomlin, R., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in second language acquisition, 16*, 183-203.

VanPatten, B., Williams, J., Rott, S., Overstreet, M. (2004). *Form-meaning connections in second language acquisition.* New Jersey: Lawrence Erlbaum Associates.

Van Hell, J. G., & Candia Mahn, A. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language learning, 47*, 507-546.

Watanabe, Y. (1997). Input, intake, and retention: Effects of increasing processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*, 287-307.

Selected Proceedings of the 2007 Second Language Research Forum

edited by Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt

Cascadilla Proceedings Project Somerville, MA 2008

Copyright information

ISBN 978-1-57473-425-6 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Martínez-Fernández, Ana. 2008. Revisiting the Involvement Load Hypothesis: Awareness, Type of Task and Type of Item. In *Selected Proceedings of the 2007 Second Language Research Forum*, ed. Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt, 210-228. Somerville, MA: Cascadilla Proceedings Project.

or:

Martínez-Fernández, Ana. 2008. Revisiting the Involvement Load Hypothesis: Awareness, Type of Task and Type of Item. In *Selected Proceedings of the 2007 Second Language Research Forum*, ed. Melissa Bowles, Rebecca Foote, Silvia Perpiñán, and Rakesh Bhatt, 210-228. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1746.