Asia-Pacific Edu Res (2021) 30(5):421–430
https://doi.org/10.1007/s40299-020-00528-8

REGULAR ARTICLE

Effects of Task Involvement Load on L2 Vocabulary Acquisition and Their Association with Language Aptitude

Yingli Yang[1] · Xiaofang Cao[2]

Published online: 1 August 2020

Abstract Vocabulary acquisition is a central component of second language learning. While there have been advancements in our understanding of the factors contributing to vocabulary acquisition in L2 students, it is still unclear how language aptitude is associated with the effects of task involvement load in this process. This study investigates the effects of task involvement load on incidental vocabulary learning and their association with language aptitude. One hundred and forty-four participants were assigned to five groups. All groups finished reading a passage and completed tasks that had different involvement loads. A pretest and two posttests adapted from Paribakht and Wesche's (TESOL Can J 11:9–29, 1993) Vocabulary Knowledge Scale (VKS) were employed to assess learners' development of the target vocabulary items. The LLAMA test (Meara, in: In_lognostics, Retrieved from February 14, 2017, from https://www.lognostics.co.uk/tools/llama, 2005) was administered to measure participants' language aptitude. We found that, in the immediate posttest, the Gap-fill (with *search*) group showed superiority over the Reading (no *need*) and Reading (with *need*) groups, while the Sentence-writing group significantly outperformed the other four groups. In the delayed posttest, both the Gap-fill (with *search*) group and the Sentence-writing group significantly outperformed the other three groups. Our study also reveals a moderate correlation between language aptitude and vocabulary learning.

Keywords Vocabulary acquisition · Task involvement load · Language aptitude · VKS

Introduction

Vocabulary acquisition is an essential aspect of any second language (L2) learning. Paribakht and Wesche (1997) argued that reading with specific word-focused tasks can lead to better vocabulary gains than a "reading only" condition. Laufer (2005) held a similar view, claiming that tasks with a lexical focus were more effective than reading alone because they engaged learners in a deeper level of processing of unfamiliar vocabulary. Based on a number of empirical studies and the depth of processing theory from cognitive psychology, Laufer and Hulstijn (2001) conceived the Involvement Load Hypothesis (ILH), proposing that tasks inducing a higher involvement load are more beneficial to vocabulary acquisition for the reason that higher involvement load entails deeper cognitive processing, which is a prerequisite for the acquisition and retention of unfamiliar vocabulary items. Task involvement thus refers to the cognitive effort the learner exerts in a word-learning task (Zou 2018).

Since the proposal of the Involvement Load Hypothesis, many researchers carried out empirical research to examine the contribution of the involvement load to vocabulary acquisition; however, these studies yielded mixed results. While some studies confirmed the Involvement Load Hypothesis (Hulstijn and Laufer 2001; Keating 2008; Kim

✉ Yingli Yang
yyluibe@126.com

Xiaofang Cao
178041243@qq.com

[1] School of International Studies, University of International Business and Economics, Room 1327, Chengxin Building, No. 10, Huixin Dongjie, Chaoyang District, Beijing 100029, China

[2] The Department of Basic Education, Wuxi Vocational Institute of Commerce, Room 408, No. 1 Teaching Building, No. 809, Qianhu Rd., Wuxi 214153, Jiangsu, China

2011), others did not support or only lent partial support to this hypothesis (Huang 2004; Wei and Wang 2011; Soleimani and Rahmanian 2015). A few studies demonstrated that individual difference factors, such as working memory and language proficiency, may mediate the effects of involvement loads on vocabulary learning (Huang 2004; Yang et al. 2017). Language aptitude, a cognitive variable hypothesized to be able to predict L2 learners' learning rate (Carroll 1981), has been shown to be associated with the effects of instructed language learning (Li 2016; Sheen 2007). However, the association of language aptitude with the effects of task involvement load on incidental vocabulary acquisition has not been examined in previous studies. Therefore, the present study seeks to investigate how task-induced involvement and language aptitude affect incidental vocabulary acquisition.

Literature Review

Involvement Load Hypothesis and Vocabulary Acquisition

Laufer and Hulstijn (2001) hypothesized that *involvement* consists of three motivational-cognitive components: *need, search,* and *evaluation*. They claimed vocabulary acquisition is conditioned upon the degree to which learners process the words. The degree of *need, search* and *evaluation* together reflects the level of cognitive processing of a given task. *Need* for knowing target words for a specific task indicates its presence. Learners' attempts to locate the meanings of words or to find a word form to express a meaning by accessing dictionaries or consulting authorities show the presence of *Search*. *Evaluation* occurs when learners compare the meanings of words or the meaning differences of a specific word in a given or self-generated context. According to the Involvement Load Hypothesis, the higher the involvement index, the more beneficial a given task will be in facilitating vocabulary acquisition. For example, a gap-fill exercise, which requires the learners to fill in the gaps with the target words, may induce less involvement load than a sentence-making or composition-writing task, which requires learners to use the target words in an appropriate context.

Hulstijn and Laufer's (2001) study directly examined the effect of task-induced involvement on the initial learning and retention of ten words in two contexts. Participants were young EFL learners from Israel and the Netherlands. Three tasks with different levels of involvement loads were employed in their study to test the Involvement Load Hypothesis: reading comprehension (moderate *need*, no *search* and no *evaluation*), gap-fill (moderate *need*, no *search*, and moderate *evaluation*), and composition writing with target words (moderate *need*, no *search* and strong *evaluation*). The results of Hulstijn and Laufer's (2001) two experiments were different. The Hebrew-English Experiment was in line with the prediction of the Involvement Load Hypothesis. There were significant differences between the composition group and the gap-fill group, and between the gap-fill group and the Reading group, which confirmed the hypothesis. However, the Dutch-English Experiment lent only partial support to the Involvement Load Hypothesis. The composition group scored higher than the gap-fill group, but the gap-fill group did not obtain significantly higher scores than the Reading group.

Kim (2011) contended that the significant differences in Hulstijn and Laufer's (2001) study may have been confounded by the time for the completion of the test for the three groups. Considering this issue, Kim designed an additional graphic organizer task for the Reading group which obliged all groups to take an equal amount of time to complete the task, thus eliminating the confounding factor of time. The results indicated that a higher level of motivational-cognitive process promoted better retention of new words. In addition, learners of two proficiency levels who completed two different tasks (composition and sentence writing) with an identical involvement load acquired a similar amount of new vocabulary, suggesting that there was no interaction effect between involvement and proficiency level on L2 vocabulary learning. Keating (2008) also provided evidence in favor of the Involvement Load Hypothesis by showing the highest retention rate in the sentence writing task, a lower rate in the reading plus fill-in task, and the lowest rate in the reading comprehension task. However, when the time factor was taken into consideration, the effects of task involvement load were attenuated.

Wei and Wang (2011) investigated the effects of task involvement load together with frequency of occurrence on the acquisition of idiomatic expressions. They found that reading with multiple choice was more effective than the Gap-fill and Sentence-writing tasks, which did not support the Involvement Load Hypothesis. Huang (2004) found that Gap-fill and Sentence-writing groups outperformed the Reading with multiple choice group, but Gap-fill had the highest gains among the three groups, which was not consistent with the Involvement Load Hypothesis. Soleimani and Rahmanian (2015) also partially supported the Involvement Load Hypothesis in that both the blank-filling group and the sentence-making group performed better than the reading comprehension group which had a lower involvement load. However, the blank-filling group and the sentence-making group had similar performance.

Most of the afore-mentioned studies seem to support the Involvement Load Hypothesis. However, mixed results remain to be explained and the contribution of each element to incidental vocabulary acquisition is still unclear. The present study aims to test the Involvement Load Hypothesis by designing five tasks which separate each component into different task conditions to further explore the extent to which each component contributes to incidental vocabulary acquisition.

Individual Difference and L2 Vocabulary Learning

In addition to task involvement load, a few studies on incidental vocabulary acquisition revealed that individual difference variables may be associated with vocabulary acquisition. Yang et al. (2017) examined the association of working memory and the effects of post-reading, word-focused activities on vocabulary acquisition. Linear regression analysis revealed that working memory was associated with the gain scores of the comprehension only and the gap-fill groups on the posttest. Huang (2004) found that high-proficiency learners benefited more from Gap-fill tasks than from tasks on the posttest, but low-proficiency learners did not demonstrate differential achievements in tasks of different involvement loads. Sarbazi (2014) confirmed that tasks of higher involvement load led to better vocabulary acquisition; however, no interaction effect between gender and involvement load on vocabulary acquisition was found. Xie et al. (2017) reviewed recent studies related to task involvement load and argued that factors such as frequency of occurrence, topic of interest of the reading material may have an impact on incidental vocabulary learning.

Although previous studies examined the influence of task-related and individual variables on incidental vocabulary acquisition, there has been few studies that investigated cognitive individual factors on vocabulary acquisition. Language aptitude reflects a set of cognitive abilities which can predict how well, compared with other individuals, one can learn a foreign language in a given amount of time and under given conditions (Carroll and Sapon 2002). A number of studies have lent support to the predictive power of language aptitude on L2 learning. For instance, Ehrman and Oxford (1995) investigated the influence of individual difference variables on adults' speaking and reading performance after a training period. They found that language aptitude was responsible for 25% of the variance, correlating most closely with participants' language performance. Wu et al. (1993) investigated the effects of psychological factors and learning strategies on Chinese students' English acquisition. Their results indicated that language aptitude had the highest predictive power, which could account for the 35.79% of the variance. In a more recent meta-analysis of language aptitude and language acquisition (including the sub-components of listening, speaking, reading and writing), Li (2016) found that language aptitude is overall significantly correlated with language learning with a medium effect size ($r = 0.49$). In terms of vocabulary learning, overall aptitude is correlated with vocabulary learning with a small effect size ($r = 0.15$), but phonetic coding ability (a subcomponent of language aptitude) was shown to have a stronger correlation with vocabulary learning ($r = 0.38$) than with other sub-skills of language learning.

Abundant empirical research has evidenced the predictive power of language aptitude on L2 learning achievement. However, to date, no empirical study has examined the association between language aptitude and the effects of tasks of different involvement loads on vocabulary acquisition. The present study aims to examine how different levels of task involvement load affect the target word learning of L2 learners with similar English proficiency levels, and the association of such learning with language aptitude.

Method

Research Questions

Our study seeks to answer the following research questions:

1. Does the level of task-induced involvement load affect L2 learners' initial vocabulary acquisition and its retention?
2. Are language aptitude (including its subcomponent) scores associated with L2 learners' initial acquisition and retention of vocabulary?

Participants

The participants were non-English majors in a vocational school ($n = 144$). All participants were young adults aged from 19 to 20. Their native language was Chinese and they had learnt English as a second language for at least 10 years. The participants were from five parallel classes in the first grade with the ratio between male and female students in each class at about two to one. From the biodata questionnaire, the participants had homogenous language learning background and language proficiency. A Kruskal–Wallis H test was adopted to analyze participants' final examination scores. There existed no significant between-group difference: $H_{(4)} = 2.371$, $p = 0.668$, which indicates that all participants had a similar English proficiency level. All five classes were randomly assigned into one of the five subgroups to complete tasks with different involvement loads.

Tasks

Five different tasks with different levels of involvement load were employed in the study as the treatment conditions.

Task one was reading with glosses irrelevant to the task for the Reading (no *need*) group. Participants were required to read the passage and complete the comprehension questions and a graphic organizer. According to Laufer and Hulstijn (2001), reading comprehension tasks with glossed words irrelevant to the questions indicates no *need*. Therefore, the task-induced involvement index of the Reading (no *need*) group was 0.

Task two was reading with glosses relevant to the task for the Reading (with *need*) group. Participants were to read the passage and complete comprehension questions related to the target words, which meant an external need for knowing the word due to the task requirement. They were also required to complete the graphic organizer. The task of the Reading (with *need*) group induced moderate *need*, no *search* and no *evaluation*, indicating an involvement index of 1.

Task three was a gap-fill exercise with no dictionary consulting for the Gap-fill (no *search*) group. Participants were required to answer comprehension questions irrelevant to the target words and fill in the appropriate words in the blanks by evaluating the eight target words and two distracters displayed in random order. In this case, according to Laufer and Huistijn (2001), moderate *evaluation* was induced. Therefore, the task of the Gap-fill (no *search*) group induced moderate *need*, no *search* and moderate *evaluation*, with an involvement index of 2.

Task four was Gap-fill (with *search*) which required participants to read the text, fill in the blanks and complete the comprehension questions. The difference between the two Gap-fill groups lied in the necessity to use dictionaries. Since the glosses were not provided, students in the Gap-fill (with *search*) group had to acquire the meaning of words by themselves. Thus, the task of the Gap-fill (with *search*) group induced moderate *need*, *search* and moderate *evaluation*, indicating an involvement index of 3.

Participants in the Sentence-writing group were required to read the reading passage (glossed), finish comprehension questions irrelevant to the target words and write original sentences with eight target words. The target words had great relevance to the requirements of the sentence writing task and students were expected to consider how to combine each given word with other additional words to form original sentences, which induced the presence of moderate *need* and strong *evaluation*. Hence, the task-induced involvement index was 3.

To control the time variable, a graphic organizer exercise was added to tasks for the Reading (no *need*) group and the Reading (with *need*) group (Kim 2011). This exercise did not focus on any of the target words, which would not affect the involvement in the two Reading groups. The differences in task-induced involvement load among the five groups are demonstrated in Table 1.

Table 1 Task involvement load index in different groups

Task	Need	Search	Evaluation	Index
Reading (no need)	0	0	0	0
Reading (with need)	1	0	0	1
Gap-fill (no search)	1	0	1	2
Gap-fill (with search)	1	1	1	3
Sentence writing	1	0	2	3

Target Words

The reading text was selected on the basis of the following four criteria: First, the length was about 500–800 words at moderate difficulty level. Second, the topic was familiar to all participants and required no particular background of this subject. Previous studies suggested that high topic familiarity may cause greater efficiency of attention allocation to reading material, thus contributing to better understanding of the text, which led to better memory performance (Ellis 2001; Nassaji 2002). Controlling the familiarity level not only made the text more appropriate for understanding, but also avoided the occurrence of another variable-reading material that is too difficult may lead to no language acquisition for any group. Third, the percentage of unfamiliar words in the reading text was controlled lower than 5% (Nation 2001). Fourth, Flesh-Kincaid readability test[1] was conducted to make sure the text was suitable for participants' English learning level.

A passage entitled *A Priceless Mother's Day Gift* with 767 words was selected as the reading material. The eight target words comprise three nouns, two verbs and three adjectives: *equanimity, infantryman, savor, reverence, condescend, beige, accusatory, gleaming*. All eight words were examined via vocabulary profiler[2] and were not present in the most frequently used 2000 words, AWL or UWL. English teachers of the participants were asked to confirm that these words were not in the vocabulary list in any current textbook. Pretest results also confirmed that participants had no prior knowledge of the meanings of these words.

[1] It can be referred to at https://en.wikipedia.org/wiki/Flesh-Kincaid_readability_tests.

[2] These words were examined via https://www4.caes.hku.hk/vocabulary/profile.htm.

Procedure

The LLAMA test and the vocabulary pre-test were conducted one week before the treatment. To measure participants' initial learning and retention of vocabulary acquisition, two vocabulary posttests were conducted: an immediate posttest right after the completion of all tasks and a delayed posttest two weeks after the treatment. The time allocated to the reading and post-reading tasks was 30 min; 10 min were allocated to each of the 3 vocabulary tests.

Testing Instruments

Language Aptitude Test

The LLAMA Language Aptitude Test (Meara 2005) was adopted to test participants' language aptitude. The aptitude test consists of four subtests: rote vocabulary learning (LLAMA_B), sound recognition (LLAMA_D), sound-symbol association (LLAMA_E), and grammatical inferencing (LLAMA_F). The total score of the four subtests signified each participant's overall aptitude score. The LLAMA test has been used in a variety of research contexts (Abrahamsson and Hyltenstam 2008; Yalcin 2012; Yilmaz 2013) and its validity and reliability has been tested (Granena 2013).

Vocabulary Tests

Three versions of the vocabulary test were employed in the experiment, including one pretest and two posttests. Each vocabulary test contained the eight target words and twelve distracters. Different distracters in each test were selected. The twenty words in each test were arranged in random order to prevent students from memorizing them mechanically from the previous tests.

The vocabulary tests were adapted from Paribakht and Wesche's (1993) Vocabulary Knowledge Scale (VKS) to measure participants' vocabulary knowledge. According to the design of present study, the test was adapted to a three-point scale test as follows:

Example item of the adapted VKS:

1. I do not know the meaning of this word.
2. I know the meaning of this word. It is __________
3. The meaning of this word is __________ and I can make a sentence using this word.

The sentence ________________________________
____

Participants were asked to indicate their level of vocabulary knowledge on the VKS of the twenty words. The scores for the three vocabulary tests ranged from 8 to 24.

Data Analysis

Data were analyzed with IBM SPSS Statistics V21.0. Kolmogorov–Smirnov and Shapiro–Wilk tests on the two posttests showed that the significance level was below 0.05 in three groups: Reading (no *need*), Reading (with *need*) and Gap-fill (no *search*). Considering that the two posttest scores were not normally distributed, nonparametric tests were carried out to analyze the data to answer the two questions. The significance level was set at 0.05. If the results revealed a significant difference, the Mann–Whitney U test was used for the post hoc test to further check differences between groups. Spearman correlation tests were carried out to analyze the correlation between aptitude and the immediate posttest scores and the delayed posttest scores. The significance level was set at 0.05.

Interrater Reliability

Two raters independently coded the entire set of the three vocabulary tests. The first rater is the second author of the present study who has over 5 years' experience of teaching English at the college level. The second rater is an experienced English teacher who has rich experience in English teaching. The second rater received extensive training for the evaluation of the vocabulary tests. The first rater rated the entire set of the data, whereas a random selection of 20% of the test papers from the pre-test, immediate post-test, and delayed post-test were given to the second rater for evaluation. Pearson correlation was conducted to examine the interrater reliability between the two raters. There was a high correlation between two raters: $r_{\text{rater1–rater2}} = 0.998$ ($p < 0.001$) for the immediate posttest and $r_{\text{rater1–rater2}} = 0.989$ ($p < 0.001$) for the delayed posttest. If large differences occurred, another two trained teachers were invited to rate and a final decision was made based on the discussion of the four teachers.

Results

Task Involvement Load and Vocabulary Acquisition

The first research question asked whether the level of task-induced involvement load would affect L2 learners' initial vocabulary acquisition and retention. The means and standard deviations of the learners' performance in the three tests are displayed in Table 2. The mean scores of the pre-test for five groups were 0.00, indicating that the

Table 2 Descriptive statistics of three vocabulary tests

Test Group[a]	N	Pre		Immediate		Delayed	
		M	SD	*M*	SD	*M*	SD
Reading (no *need*) (0)	30	0.00	0.00	11.60	1.77	9.30	1.02
Reading (with *need*) (1)	30	0.00	0.00	11.67	1.77	9.27	1.17
Gap-fill (no *search*) (2)	32	0.00	0.00	12.34	2.04	9.66	1.54
Gap-fill (with *search*) (3)	26	0.00	0.00	13.92	2.77	11.15	2.48
Sentence-writing (3)	26	0.00	0.00	15.12	2.55	11.73	2.44

[a] The figure in the parentheses following the task type is the involvement index

groups did not have any prior knowledge of the meaning of the target vocabulary items.

Because the vocabulary scores of the immediate posttest did not meet the normality assumption, the scores were submitted to the nonparametric Kruskal–Wallis *H* test with the immediate posttest scores as the test variable and the group as the grouping variable. The Kruskal–Wallis test results revealed that there existed significant differences between the five task groups: $H_{(4)} = 39.310$, $p < 0.05$. Mann–Whitney *U* post hoc comparisons were adopted to explore where the significant differences existed (Qin and Bi 2015). To adjust for multiple comparisons, the significance level was set at 0.005, according to the Bonferroni correction. Significant differences were found between the following 6 pairs of groups (see Table 3): the Sentence-writing group and the Reading (no *need*) group ($u = 94.00$, $z = -4.90$, $p < 0.001$), with a large effect size ($r = 0.644$) (Cohen 1988); the Sentence writing group and the Reading (with *need*) group ($u = 95.000$, $z = -4.881$, $p < 0.001$), with a large effect ($r = 0.653$); the Sentence writing group and the Gap-fill (no *search*) group ($u = 154.000$, $z = -4.134$, $p < 0.001$), with a large effect ($r = 0.543$); the Sentence-writing group and the Gap-fill (with *search*) group ($u = 66.500$, $z = -4.997$, $p < 0.001$), with a large effect ($r = 0.695$); the Gap-fill (with *search*) group and the Reading (no *need*) group ($u = 186.000$, $z = -3.379$, $p = 0.001$), with a medium effect size ($r = 0.452$); the Gap-fill (with *search*) group and the Reading (with *need*) group ($u = 189.50$, $z = -3.321$, $p = 0.001$), with a medium effect ($r = 0.444$).

The results lend partial support to the ILH in that the Sentence-writing group, with an involvement load index of 3, outperformed all the other groups, and the Gap-fill (with *search*) group, with identical involvement of 3, also achieved significantly higher gains than the two reading groups with lower involvement loads. However, there also existed a significant difference between the two groups with the identical index, the Sentence-writing group outperforming the Gap-fill (with *search*) group, which we discuss below in the Discussion. Additionally, no significant difference was found between the Reading (no *need*) group and the Reading (with *need*) group, between the Reading (no *need*) group and the Gap-fill (no *search*) group, between the Reading (with *need*) group and the

Table 3 Pairwise comparisons of different groups in the immediate posttest

Test groups[a]	Sig.	Effect size (*r*)
Sentence writing (3) > Reading (no *need*) (0)	0.000	0.644
Sentence writing (3) > Reading (with *need*) (1)	0.000	0.653
Sentence writing (3) > Gap-fill (no *search*) (2)	0.000	0.543
Sentence writing (3) > Gap-fill (with *search*) (3)	0.000	0.695
Gap-fill (with *search*) (3) > Reading (no *need*) (0)	0.001	0.452
Gap-fill (with *search*) (3) > Reading (with *need*) (1)	0.001	0.444

[a] The figures in the parentheses following the task type indicate the involvement index

Table 4 Pairwise comparisons of different groups in the delayed posttest

Test groups[a]	Sig.	Effect size (*r*)
Sentence writing (3) > Reading (no *need*) (0)	0.000	0.639
Sentence writing (3) > Reading (with *need*) (1)	0.000	0.629
Sentence writing (3) > Gap-fill (no *search*) (2)	0.000	0.508
Gap-fill (with *search*) (3) > Reading (no *need*) (0)	0.000	0.470
Gap-fill (with *search*) (3) > Reading (with *need*) (1)	0.000	0.480
Gap-fill (with *search*) (3) > Gap-fill (no *search*) (2)	0.004	0.375

[a] The figures in the parentheses following the task type indicate the involvement index.

Gap-fill (no *search*) group, or between the Gap-fill (with *search*) group and the Gap-fill (no *search*) group ($p > 0.005$).

The Kruskal–Wallis *H* analysis revealed that there were statistically significant differences among the five groups in the delayed posttest, $H_{(4)} = 37.836$, $p = 0.000$. Mann–Whitney *U* post hoc results indicated significant differences among the 6 pairs (see Table 4): the Sentence writing group and the Reading (no *need*) group ($u = 103.000$, $z = -4.784$, $p < 0.001$), with a large effect size ($r = 0.639$) (Cohen 1988); the Sentence-writing group and the Reading (with *need*) group ($u = 107.500$, $z = -4.709$, $p < 0.001$), with a large effect size ($r = 0.629$); the Sentence-writing group and the Gap-fill (no *search*) group ($u = 172.000$, $z = -3.872$, $p < 0.001$), with a large effect size ($r = 0.508$); the Gap-fill (with search) group and the Reading (no *need*) group ($u = 179.500$, $z = -3.518$, $p < 0.001$), with a medium effect size ($r = 0.470$); the Gap-fill (with *search*) group and the Reading (with *need*) group ($u = 175.000$, $z = -3.592$, $p < 0.001$), with a medium effect size ($r = 0.480$); the Gap-fill (with *search*) group and the Gap-fill (no *search*) group ($u = 236.500$, $z = -2.853$, $p = 0.004$), with a medium effect size ($r = 0.375$).

The results of this analysis of the posttest data offer partial evidence in support of the Involvement Load Hypothesis. In the delayed posttest, the Sentence-writing group and the Gap-fill (with *search*) group, both with the highest involvement loads of 3, significantly outscored the other three groups, but there was no significant difference between the two groups. No significant difference was found between the Reading (no *need*) group and the Reading (with *need*) group, the Reading (no *need*) group and the Gap-fill (no *search*) group or the Reading (with *need*) group and the Gap-fill (no *search*) group ($p > 0.005$).

Language Aptitude and Vocabulary Acquisition

The second research question regarded the association between language aptitude and L2 learners' vocabulary acquisition. As the vocabulary scores of the participants did not meet the normality assumption for parametric tests, Spearman correlation analyses were conducted to investigate the correlation between language aptitude and participants' vocabulary achievements. The significance level was set to 0.05. We found significant correlations between aptitude and the immediate posttest vocabulary achievements of the Reading (with *need*) group ($r = 0.422$, $p = 0.020$), and the Gap-fill (with *search*) group ($r = 0.450$, $p = 0.021$) with medium effect sizes (Cohen 1988). A further look at the correlation coefficients between language aptitude (including its subcomponents) and the immediate posttest scores revealed that LLAMA_B significantly correlated with scores in the Gap-fill (with *search*) group ($r = 0.398$, $p = 0.044$), and the Gap-fill (no *search*) group ($r = 0.419$, $p = 0.017$). LLAMA_F significantly correlated with scores in the Reading (no *need*) group ($r = 0.422$, $p = 0.020$) and the Gap-fill (no *search*) group ($r = 0.349$, $p = 0.050$).

The delayed posttest results revealed a statistical significance between aptitude and the Gap-fill (no *search*) group ($r = 0.408$, $p = 0.020$), while the correlation between language aptitude and the Gap-fill (with *search*) group was approaching significance ($r = 0.387$, $p = 0.051$), with medium effect sizes. LLAMA_F significantly correlated with delayed posttest scores in the Gap-fill (no *search*) group ($r = 0.437$, $p = 0.012$) and the Reading (no *need*) group ($r = 0.474$, $p = 0.008$).

Discussion

We found that the Sentence-writing group has an overall advantage over all the other groups in the immediate posttest, followed by the Gap-fill (with *search*) group which outperformed the two Reading groups. These results are in line with the results of previous studies that showed that tasks with higher involvement loads had an advantage over tasks with lower involvement loads (Keating 2008; Kim 2011; Laufer and Hulstijn 2001; Yang et al. 2017).

Interestingly, our study revealed that, in the immediate posttest, the two tasks with an identical involvement load of 3 yielded significantly different scores in terms of initial vocabulary acquisition, with the Sentence-writing group showing a significant advantage over the Gap-fill (with *search*) group, which is different from the results reported in the study by Kim (2011). In her study, two tasks (Sentence-writing and Composition) with an identical involvement load (moderate *need*, no *search* and strong *evaluation*) led to similar gains on vocabulary acquisition. The possible reason for the discrepancy between our study and hers is that the two groups with the identical involvement load in the present study differ in terms of the three components entailed. The Sentence-writing group induces moderate *need*, no *search* and strong *evaluation*, whereas the Gap-fill group involves moderate *need*, *search* and moderate *evaluation*. Consistent with the results reported in Yang et al. (2017) and Laufer (2003), our present study shows that strong *evaluation* may be "the most influential factor for learner's initial vocabulary acquisition" (Kim 2011, p.125), at least in the short term. Even though the amount of the involvement load is considered as a whole, the separate component (*need, search,* and *evaluation*) may weigh differently for vocabulary learning. Huang et al.'s (2012) meta-analytic study also showed that the mean effect size of the Sentence writing group (0.94) is greater than that of the Gap-fill group (0.81). The underlying reason might be that writing original sentences involves the production of the target word in a new context, which requires deeper cognitive processing and facilitates memorization of new vocabulary items. In addition, writing original sentences requires learners to combine different words to form an original context, while filling gaps only allows learners to process lexical information in a given context. As Joe's (1995, 1998) studies revealed, using new words in an original context results in better memorization than utilizing them in non-original contexts.

In both posttests, the Reading (no *need*) group and the Reading (with *need*) group made similar gains, which implied that the *need* component may not be a significant factor in differentiating vocabulary acquisition scores between the two groups. By contrast, when considering the Gap-fill (with *search*) group and the Gap-fill (no *search*) group, one may conclude that, with the other two components being equal, the significant difference in the delayed posttest was caused by the *search* component. From the above pairwise comparisons, it can be inferred that despite two factors being equally marked as 1, the impact of *search* on vocabulary incidental learning was greater than that of *need*. Skehan (1989) argued that *need* can only indirectly facilitate vocabulary learning, which would not effectively predict vocabulary scores. There have been quite a few researchers asserting that consulting dictionaries facilitates incidental vocabulary learning better than marginal glosses (Hulstijn et al. 1996; Laufer 2000). Keating (2008) contended that "looking up unknown words in a dictionary while reading (*search*) draws learners' attention to form in a way that is much more overt than when words are glossed in the margin (no *search*)" (p. 368).

Regarding the association between language aptitude and vocabulary acquisition, our results generally confirm Li's (2016) findings that aptitude, as a whole, weakly correlates with learners' vocabulary scores, but phonetic coding and language analytic ability show a moderate correlation with vocabulary learning. The reason for these relatively less than robust correlations may be that, in our study, participants were assigned to different reading tasks and their attention was not intentionally directed to the vocabulary embedded in the reading material; vocabulary acquisition was just a by-product of reading activities. Reber et al. (1991) claimed that aptitude may only be predictive for language learning in explicit conditions. And Li (2015) presented a similar view that aptitude was more strongly correlated with explicit treatments than implicit treatments. However, our results did show that LLAMA_B, a measure that assesses learners' ability to learn new vocabulary items (Granena 2013), significantly correlated with vocabulary gains in the Gap-fill (no *search*) group and Gap-fill (with *search*) group. This is possibly because, while learners evaluate the meaning of different words and fill in these words in appropriate contexts, the mechanism of associative learning ability and analytical ability entailed in LLAMA_B was accessed. LLAMA_F, which measures inductive language learning ability, associated with vocabulary scores in the Reading (no *need*) and Gap-fill (no *search*) groups in both posttests. It is possible that in tasks where the element *search* is absent, strong analytical skills are important in assisting learners in noticing the mapping between form and meaning of the vocabulary items, which in turn leads to better acquisition and retention of vocabulary items.

Conclusion

Our study set out to investigate the role of task involvement load and language aptitude on vocabulary learning. The results partially support the Task Involvement Load Hypothesis proposed by Laufer and Hulstijn (2001). Generally, tasks with higher involvement loads were more beneficial for L2 learners' vocabulary acquisition than those with lower involvement loads. The results of pairwise comparisons indicated that strong *evaluation* was the most important factor for vocabulary acquisition, and *search* was more powerful than *need*, concerning their effects on vocabulary learning, which is consistent with Rott's (2005) argument that searching for meaning and evaluation for proper words in a context contribute to better vocabulary acquisition.

Regarding the predictive power of aptitude on vocabulary scores in posttests, LLAMA_B and LLAMA_F scores were shown to have moderate correlations with vocabulary scores in the two posttests. These two subcomponents of the LLAMA test, which tap into learners' associative learning ability and analytical ability, were significant predictors of vocabulary acquisition.

To a large extent, our findings confirm the necessity of combining reading with word-focused, post-reading exercises to assist learners' vocabulary acquisition. Instructors can design reading tasks with high involvement loads to help students improve the efficiency of vocabulary acquisition. Tasks with strong *evaluation* and *search*, such as summary, composition and sentence writing tasks, and allowing learners to consult dictionaries for the meanings of unfamiliar words, are recommended. In addition, the significant mediating role of aptitude on vocabulary acquisition suggested that activities that are designed to train learners' associative learning abilities and analytical abilities may also facilitate vocabulary acquisition. Learners may also benefit from tasks that induce higher involvement loads since these tasks require them to process the words at a deeper level and thus lead to better memorization and retention.

There are some limitations in our study. First, the number of participants in each group was relatively limited. Moreover, the delayed posttest was administered only two weeks after the treatment. In order to gain a better and more profound understanding of the long-term effects of task involvement load, a study with a longitudinal design is needed. Finally, additional empirical studies are needed to further investigate the relationship between language aptitude together with other individual difference variables and vocabulary acquisition under different instructional conditions.

Acknowledgements This study is funded by the Ministry of Education in China (MOE) Project of Humanities and Social Sciences Research [grant No. 20YJA740052]. The authors would like to thank Dr. Alex Sheftel for proofreading the manuscript.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*, 481–509.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–117). Rowley, MA: Newbury House.

Carroll, J. B., & Sapon, S. M. (2002). *Modern language aptitude test manual*. Bethesda, MD: Second Language Testing Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ehrman, M., & Oxford, R. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79*, 67–89.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). Cambridge: Cambridge University Press.

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). Amsterdam: John Benjamins Publishing Company.

Huang, S., Eslami, Z., & Wilson, V. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A mete-analytic study. *The Modern Language Journal, 96*(4), 544–557.

Huang, Y. (2004). An empirical study on the Task Involvement Load Hypothesis: Effects of reading tasks on Chinese students' vocabulary acquisition. *Modern Foreign Languages, 27*(4), 386–394.

Hulstijn, J. H., Hollander, B., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80*, 327–339.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*, 539–588.

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

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*, 357–377.

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

Kim, Y. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 61*(Suppl. 1), 100–140.

Laufer, B. (2000). Electronic dictionaries and incidental vocabulary acquisition: Does technology make a difference? In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *EURALEX* (pp. 849–854). Stuttgart: Stuttgart University Press.

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

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook, 5*, 223–250.

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

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analysis review of five decades of research. *Applied Linguistics, 36*(3), 385–408.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition, 38*(4), 801–842.

Meara, P. (2005). Llama language aptitude tests. *In_lognostics*. Retrieved from February 14, 2017, from https://www.lognostics.co.uk/tools/llama.

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

Nassaji, H. (2002). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language Learning, 52*(2), 439–482.

Paribakht, T. S., & Wesche, M. (1993). The relationship between reading comprehension and second language development in a comprehension-based ESL program. *TESOL Canada Journal, 11*, 9–29.

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

Qin, X., & Bi, J. (2015). *Quantitative approaches and quantitative data analysis in L2 research*. Beijing: Foreign Language Teaching and Research Press.

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*(5), 888–896.

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.

Sarbazi, M. (2014). Involvement Load Hypothesis: Recalling unfamiliar words meaning by adults across genders. *Procedia - Social and Behavioral Sciences, 98*, 1686–1692.

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 301–322). New York, NY: Oxford University Press.

Skehan, P. (1989). *Individual differences in second language acquisition*. London: Edward Anold.

Soleimani, H., & Rahmanian, M. (2015). Vocabulary Acquisition and Task Effectiveness in Involvement Load Hypothesis: A case in Iran. *International Journal of Applied Linguistics & English Literature, 4*(5), 198–205.

Wei, M., & Wang, L. (2011). Effects of task type on university students' acquisition of idiomatic expressions: Reexamining the Task Involvement Load Hypothesis. *Modern Foreign Languages, 34*(4), 372–380.

Wu, Y., Liu, R., & Jeffrey, P. (1993). Reports on Chinese undergraduate students' English learning ability. *Foreign Language Teaching and Research, 1*, 36–46.

Xie, H., Zou, D., Wang, F., & Wong, T. (2017). A review on recent development of the Involvement Load Hypothesis. In *International conference on blended Learning ICBL 2017: Blended learning. New challenges and innovative practices* (pp. 447–452). Cham: Springer.

Yalcin, S. (2012). *Individual differences and the learning of two grammatical features with Turkish learners of English*. Unpublished Doctoral Dissertation. University of Toronto.

Yang, Y., Shintani, N., Li, S., & Zhang, Y. (2017). The effectiveness of post-reading word-focused activities and their association with working memory. *System*, *70*, 38–49.

Yilmaz, Y. (2013). The role of working memory capacity and language analytical ability in the effectiveness of explicit correction and recasts. *Applied Linguistics, 34*(2), 344–368.

Zou, D. (2018). Vocabulary acquisition through cloze exercises, and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research, 21*(1), 54–75.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.