

Comparing Input and Output Tasks in EFL Learners' Vocabulary Acquisition

Gui Bao¹

Nanjing Tech University, China

Abstract

Task-based vocabulary learning has recently attracted lots of attention in the field of second language (L2) vocabulary acquisition. This article compares the effects of input and output tasks on English as a Foreign Language (EFL) learners' acquisition of vocabulary knowledge. Four intact classes of EFL learners were randomly assigned to one of four tasks of learning 18 target words through sentence reading exercises, i.e., matching, definition, choice, and combining. The definition task was found to be more effective than the other tasks in EFL vocabulary knowledge, irrespective of EFL proficiency. At each level of EFL proficiency, the matching and combining tasks performed equally well. The choice task had a small advantage over the matching and combining tasks at the learners' low EFL proficiency level, but not at the intermediate or high EFL proficiency level. Differential processing and division of attention may well affect task effectiveness.

Keywords: input task, output task, EFL vocabulary knowledge, differential processing, division of attention

Introduction

It is generally acknowledged that second language (L2) reading with related word-focused tasks (i.e., reading plus) is more conducive to vocabulary acquisition than L2 reading without such tasks (i.e., reading only) (Laufer & Girsai, 2008; Paribakht & Wesche, 1997; Wesche & Paribakht, 2000). Word-focused tasks through reading arouse the learners' attention to new words, thereby increasing the chances that the words will be retained. Such attention may not be necessarily evoked during a "reading only" task, whose purpose is to gain an overall understanding of the text (Hill & Laufer, 2003, p. 90). L2 researchers and teachers have developed a variety of tasks to ascertain whether some tasks are more effective than others in improving vocabulary knowledge (Barcroft, 2002; Golonka et al., 2015; Huang & Lin, 2014; Joe, 1998). One important question arising from this line of research is how to theorize task effectiveness. Of most recent theoretical interest is the Involvement Load Hypothesis (ILH) (Laufer & Hulstijn, 2001).

The ILH accrues from Laufer and Hulstijn's (2001) concern for the failure of the levels of processing (LOP) framework to provide clear definitions of notions like "depth of processing" and "degree of elaboration" (Craik & Lockhart, 1972; Craik & Tulving, 1975), and for the possible effects of human and social-cultural factors like motivation on information processing. According to the ILH, the construct of involvement is defined as consisting of three concrete task-related components, i.e., need (motivational dimension), search and evaluation (cognitive or information processing dimensions). The involvement load is thus determined by the presence and strength of each component. The ILH predicts that, the greater the task-induced involvement loads, the more likely the word will be learned. A number of researchers (e.g., Folse, 2006; Hulstijn & Laufer, 2001; Keating, 2008) designed various L2 vocabulary learning tasks to test the ILH. Nevertheless, the hypothesis was supported only partially, suggesting that task effectiveness is not always revealed by the involvement load a task generates.

One of the crucial issues is the dependence of task effectiveness on the input- or output-orientation. The ILH hypothesizes that word retention is contingent upon nothing but

E-mail: boggy2008@126.com. School of Foreign Languages and Literature, Nanjing Tech University, No. 30, South Puzhu Road, Pukou District, Nanjing 211816, China.

the involvement load a task generates, be it input- or output-oriented. This suggests that input and output tasks would be equally effective if they generated identical involvement loads. Although numerous studies have addressed the ILH, no due attention has been paid to this suggestion. Accordingly, the current study seeks to examine the effects of several input and output tasks with the same involvement loads on English as a Foreign Language (EFL) vocabulary learning.

Background Literature

The ILH focuses essentially on semantic spread (i.e., the degree to which semantic information is enriched in processing new words), ascribing the learner's retention of hitherto unfamiliar L2 words (or vocabulary items) to the synergism of the three components of task-induced involvement, i.e., need (N), search (S), and evaluation (E). According to Laufer and Hulstijn (2001), "need" is a drive to meet the task demands; it may be absent (-N) when the task is not relevant to the new words, or may have a moderate presence (+N) if it is imposed by an external agent, for instance, if L2 reading comprehension questions are relevant to the new words glossed in the text, or even a strong presence (++N) if it is intrinsically motivated by the learner per se, for instance, in a composition where the learner decides to consult a bilingual dictionary for the unknown equivalents of certain L1 concepts. "Search" is an attempt to find the form or meaning of an unknown word; it may be either absent (-S) if this attempt is not made, or present (+S) if it is. "Evaluation" involves a decision about the meaning or usage of a new word. It may be absent (-E) when the task is not relevant to the new words, or may have a moderate presence (+E) when the task entails recognition of differences among words (as in a fill-in task), or among several senses of a word in the specific context, or even a strong presence (++E) when the task entails the use of a new word in an original text. The involvement load is thus operationalized as the involvement index, with the absence of a component marked as 0, moderate presence as 1, and strong presence as 2. The ILH assumes that each component of involvement carries equal weight in relation to word retention, and that the involvement indexes can be simply added to represent the degree of overall involvement.

Most empirical research regarding the ILH has been conducted through learning words in reading passages or single sentences. In both sizes of context, however, mixed evidence is found for the ILH.

The first mixed evidence was derived from Hulstijn and Laufer (2001), who compared three tasks through passage reading: reading [+N, -S, -E], reading plus fill-in [+N, -S, +E], and composition writing [+N, -S, ++E]. The writing task was found to produce significantly better retention of the target word meanings than both the reading and reading plus fill-in tasks in two experiments, whereas the reading plus fill-in task resulted in better retention than the reading task in one experiment but not in the other, thus largely but not fully supporting the ILH. The majority of follow-up studies also lent partial support to the ILH (in passage contexts: Eckerth & Tavakoli, 2012; Keating, 2008; Kim, 2008; Laufer & Rozovski-Roitblat, 2011; in sentence contexts: Bao, 2015; Folse, 2006; Webb, 2005; Webb & Kagimoto, 2009), with those in full support of the ILH being in the minority (e.g., in passage contexts: Laufer & Girsai, 2008; Min, 2008; in sentence contexts and on overall word recall: Pichette, de Serres, & Lafontaine, 2012). Among the factors affecting word retention are time and type of vocabulary knowledge measurement (e.g., Keating, 2008; Kim, 2008; Webb, 2005), word encounter frequency or retrieval frequency (e.g., Eckerth & Tavakoli, 2012; Folse, 2006; Laufer & Rozovski-Roitblat, 2011), task design (e.g., Eckerth & Tavakoli, 2012; Keating, 2008; Kim, 2008), and word characteristics (e.g., Pichette, de Serres, & Lafontaine, 2012).

To summarize, previous empirical studies regarding the ILH either involved comparing input and output tasks with different involvement loads (e.g., Hulstijn & Laufer, 2001; Webb, 2005), or output tasks inducing identical or different loads (e.g., Bao, 2015; Folse, 2006; Kim, 2008). An interesting question is, would the input- and output-oriented tasks be equally effective for L2 vocabulary learning if the task-induced involvement loads were identical? Although the ILH predicts that higher task-induced involvement will lead to better retention of an unknown word, regardless of whether the task is input- or output-oriented (Laufer & Hulstijn, 2001, p. 20), it is possible that input and output tasks, which entail two qualitatively different modes of

processing, may contribute to L2 vocabulary learning in ways unpredicted by the ILH. Therefore, the problem of concern is how the input and output tasks compare in L2 vocabulary learning if the task-induced involvement loads are held constant.

Another question of pedagogical interest is whether task effectiveness regarding vocabulary learning is related to L2 learner factors like L2 proficiency. If the positive effect of a task as compared to another task persists across L2 proficiency levels, this task would apply to the normal classroom setting where a regular class consists of L2 learners with mixed L2 proficiency levels. In one of the two experiments on English as a Second Language (ESL) learners with two levels of ESL proficiency, Kim (2008) compared two tasks, which were hypothesized to have the same level of task-induced involvement (composition writing (+N, -S, ++E), sentence writing (+N, -S, ++E)). In initial word learning, there was neither a significant main effect for either task type or ESL proficiency nor a significant two-way interaction. However, it remains to be found whether the independence of task effectiveness from language proficiency is generalizable to tasks other than composition writing and sentence writing.

Thus, this study addressed the following two questions:

1. How does task type (i.e., matching, choice, definition, combining) affect EFL learners' vocabulary knowledge?
2. Do the task type effects on EFL learners' vocabulary knowledge vary with EFL learners' proficiency?

Method

Research Design

This study was an examination of how word-focused tasks would affect EFL learners' initial word learning, employing a pretest-posttest experimental design. Task type, a between-subjects factor, had four levels, consisting of two input tasks (i.e., matching and choice) and two output tasks (i.e., definition and combining). These tasks were assumed to have identical involvement loads, but differed in the input-output orientation, presence of a target word, or both (see the Input and output tasks section). Initial word learning was measured by an immediate posttest of EFL learners' passive recall and use of the newly learned target words. Unlike the posttest, however, the pretest was a test of EFL proficiency. The whole experiment was conducted in class. During the experiment, all classes were asked to do the same sentence reading exercises involving the target words, but to complete a different word-focused task assigned to them within the allocated time limits. A short time after task completion, all the classes were asked to take the test of the target words.

Allocating different time limits to different tasks was due to the relationship between a task and task time. That is, all else being equal, tasks of different types tend to require different lengths of time to complete. Like a number of previous studies (Folse, 2006; Hulstijn & Laufer, 2001; Keating, 2008), this study considered task time to be an internal part of a task.

Participants

The participants were 167 first year intermediate EFL learners from different non-English specialties at one Chinese university. They are mostly female (151 females vs. 16 males), and their ages ranged from 16 to 22 years old ($M = 18.51$, $SD = 0.86$). These participants had learned English at school for at least six years before they went to university. They had to continue to learn English as a compulsory course at university for the first two years. They were taught English in one of the four randomly assigned intact classes for 4 hr per week.

Each intact class was randomly assigned to one of the four vocabulary learning tasks. One month before the experiment, all four classes of EFL learners took an EFL proficiency test with a full score of 100. The descriptive statistics for their EFL proficiency are displayed in Table 1, including sample size (n), mean (M), standard deviation (SD) and Shapiro-Wilk Test.

Table 1
Descriptive statistics for EFL proficiency across task type

	<i>n</i>	<i>M</i>	<i>SD</i>	Shapiro-Wilk Test	
				<i>W</i>	<i>p</i>
Matching	42	67.39	12.14	0.93*	0.018
Choice	44	67.14	12.09	0.92**	0.004
Definition	44	66.67	7.96	0.91**	0.002
Combining	37	67.91	7.93	0.98	0.701

* and ** indicate significance at the 0.05 and 0.01 levels, respectively.

As Table 1 indicates, all four groups had similar EFL proficiency scores when they took the test one month previously. For the combining task, the data are approximately normally distributed, $w = 0.98$, $p > 0.05$, but for the other tasks, the assumption of normality was violated, $p < 0.05$ or 0.01. Levene's test found that the assumption of homogeneity of variances among the groups was violated, $F(3,163) = 2.97$, $p = 0.033 < 0.05$.

A robust ANOVA using 20% trimmed means (\bar{X}) and 20% Winsorized variances (s^2)

found no statistically significant difference, $F(3, 55) = 0.15$, $p = 0.928 > 0.05$, so the four groups were considered equivalent in EFL proficiency.

Target Words and Sentence Reading Exercises

The participants in this study were at the same EFL proficiency level as those in Bao (2015), so the same 18 target words used in his study were also used in this study. These target words consisted of an equal number of nouns, verbs and adjectives. The nouns were *accessory*, *bristle*, *cabaret*, *fracture*, *gimmick*, *palette*, the verbs *allege*, *haunt*, *mumble*, *scrub*, *shudder*, *strangle*, and the adjectives *candid*, *cavalier*, *devious*, *egalitarian*, *erratic*, *malign*.

EFL reading exercises, serving as language input, comprised 18 semantically disconnected reading sentences, each of which involved a target word. These sentences were the same as in Bao (2015). The reading sentences were randomly divided into three sets of six sentences each. Immediately after each reading sentence, the gloss of a target word was given in the brackets, including its L1 (Mandarin Chinese) translations, part of speech and inflection. Below each reading sentence was a 4-point self-report scale (1 = *not understood*, 2 = *partially understood*, 3 = *largely understood*, 4 = *totally understood*). The purpose of the scale was to encourage the participants to read each sentence carefully lest they should jump to the input or output task assigned to them.

Input and Output Tasks

In both the matching and definition tasks, several definitions/descriptions as one of the three sets were presented on the right, and on the left were the target words or the blanks to be filled in with the target words. The matching task was input-oriented, while the definition task output-oriented. The participants on the matching task were instructed to draw a line linking each target word to one definition/description. The purpose of giving more definitions/descriptions than the target words was to reduce reliance on guessing. The definition task omitted the distracters used in the matching task, and replaced the target words with blanks. The participants on this task were asked to fill in each blank with the target word matching each definition/description.

For both the choice and combining tasks, the 18 sentences were evenly divided into three sets of six sentences each. Each sentence was segmented into five word strings. In both tasks, the participants were asked to rearrange the word strings into a grammatically correct sentence. The choice task was input-oriented while the combining task output-oriented. In the choice task, the word strings, numbered 1 to 5, were randomly presented in each item stem, followed by four sequences numbered A to D for the participants to choose from, but only one sequence was grammatically correct. In the combining task, however, the numbers 1 to 5 in each item stem were omitted, and the adjacent word strings were separated with semicolons. The participants were asked to write down the correct sentence order in the line below the item.

All the tasks induced moderate need (+N), since the need was imposed by the task instructions. They demanded no search for the target words (-S), since the glosses of these words were given. Both the matching and definition tasks induced moderate evaluation (+), as evaluation “entails recognizing differences between words” (Hulstijn & Laufer, 2001, p. 544). Both the choice and combining tasks also induced moderate evaluation (+), as evaluation entailed comparing word strings and deciding how the target word or word string combined with others into a given sentence or text rather than requiring a decision as to “how additional words will combine with the new word in an original (as opposed to given) sentence or text” (ibid., p. 544). Therefore, all the tasks in this study were assigned an involvement load index of 2.

Vocabulary Knowledge Test Instrument

The vocabulary knowledge test instrument was an immediate posttest of the target words. It was adapted from Min’s (2008) 4-point version of the five-point Vocabulary Knowledge Scale (VKS), which was developed by Paribakht and Wesche (1997). Min (2008) used the unknown/known word dichotomy to distinguish the first two statements (the unknown word category; Categories I and II) from the second two statements (the known word category; Categories III and IV). According to Paribakht and Wesche (1997), the first two categories were focused on form and meaning recognition (in a self-report form), whereas the last two were intended to measure meaning recall and comprehension (in a production task). This study employed Categories III and IV to measure the learner’s demonstrable vocabulary knowledge of each target word.

For the participants’ understanding, the test instructions were given in Mandarin Chinese, their native language. The participants were presented with a list of the target words and instructed to indicate their levels of knowledge for each. For Category III, the participants were asked to write down the English definition or Chinese equivalent of each target word. Those who could complete Category III should proceed to Category IV, where they had to write a meaningful sentence with each target word.

Procedures

One week before the experiment, the researcher trained four EFL teachers on how to handle the materials, and answered questions about the instructions. The experiment was carried out during the regular class periods. Since the teaching schedules were not exactly the same, each teacher was allowed to administer the experiment in their own class sessions on a separate day of the same week.

Different time limits were set for each task (i.e., each vocabulary learning task plus sentence reading exercises). Following Bao (2015), the time limits of 15 and 30 min were set for the definition and combining tasks, respectively. Since the matching and definition tasks were similar in design format, and so were the choice and combining tasks, the matching and choice tasks were given the same time limits as the definition and combining tasks, respectively.

At the beginning of the experiment, all the participants were instructed to do the reading exercises first, followed by the assigned word-focused tasks. While performing the vocabulary learning tasks, the participants were allowed to refer to the reading materials for better understanding of the target words. 10 min after the teachers collected all the materials, the participants were unexpectedly given a 20-min posttest of the target words. All the test papers were collected by the teachers when the time was up.

Scoring and Data Analyses

This study dichotomized each participant’s responses to Categories III and IV. For Category III, a score of 1 was awarded for a correct synonym or translation of the target word, but a score of 0 for no attempted response or an incorrect synonym or translation. No penalty was given for the wrong Chinese characters in the translation or replacing the correct Chinese characters with Chinese phonetic symbols as long as the correct meaning could be reasonably guessed. Regarding Category IV, a score of 1 was awarded if both the meaning and grammatical usage of the target word were correct in the sentence, regardless of errors elsewhere in the sentence, but a score of 0 was given otherwise.

Two experienced EFL teachers were trained to rate the participants' Categories III and IV for each target word independently. The inter-rater agreement was 100%, as all disagreements were discussed between the two teachers until consensus was reached. Each participant's vocabulary knowledge was represented by the cumulative scores of Categories III and IV.

To address the first research question, one robust one-way ANOVA was conducted on EFL vocabulary knowledge with task type as the between-subjects factor, followed by six post hoc linear contrasts (Hochberg's method was used to control the family-wise error rate). To address the second research question, one robust ANCOVA³ was conducted on EFL vocabulary knowledge to test the task type effects at each of three design points of EFL proficiency (covariate). R 3.5.1 was run for all data analyses. All functions for robust estimates came from Wilcox (2017). The statistical significance level was set at $\alpha = 0.05$. One robust explanatory measure of effect size, called d_w , was used. Under normality and homogeneity, $d = 0.20, 0.50$ and 0.80 represent small, medium and large effect sizes, respectively (Cohen, 1988). In this study, this criterion for the magnitudes of effect sizes was followed, and a medium effect size or beyond was deemed important.

Results

Task Type and EFL Vocabulary Knowledge

This section graphically compares the input and output tasks in vocabulary learning. Figure 1 displays the data patterns of EFL vocabulary knowledge for each task.

Figure 1
Violin plots of vocabulary knowledge vs. task type

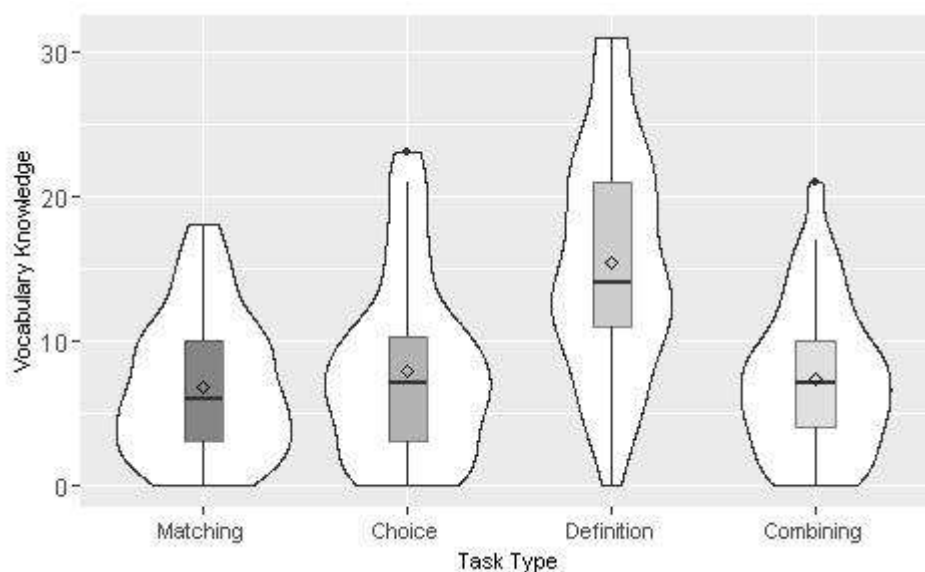


Figure 1 consists of four violin plots, i.e., kernel density plots superimposed in a mirror-image fashion over boxplots. Here, the boxes range from the lower to the upper quartile, the solid black line and the diamond in each box represent the median and mean respectively, and the black dots are outliers. Several features are discernible in this figure. To start with, the data distribution for each task is basically unimodal and right-skewed to a different extent. Furthermore, the data distribution for each task is somewhat platykurtic, especially for the definition task. Thirdly, one outlier is present for both the choice and combining tasks, but not for the other tasks. Finally, a comparison of the means and medians shows that the definition task performs remarkably better than the other three tasks, which perform almost equally well.

Effects of Task Type on EFL Vocabulary Knowledge

Table 2 reports the descriptive statistics for the vocabulary knowledge data, including sample size (n), 20% trimmed mean (\bar{X}), 20% Winsorized variance (s^2), and 95% CI.

Table 2

Descriptive statistics for EFL vocabulary knowledge across task type

	<i>n</i>	\bar{x}	s^2	95% CI
Matching	42	6.31	10.65	4.58-8.04
Choice	44	7	14.37	5.05-8.95
Definition	44	15.18	29.03	12.40-17.96
Combining	37	6.83	13.93	4.71-8.95

As Table 2 shows, the definition task fares best in vocabulary learning, and its 95% CI of the trimmed mean does not overlap with any other task's, suggesting a significant difference between the definition and any of the other tasks. Much overlap in 95% CIs suggests no significant difference among the matching, choice and combining tasks.

A one-way robust ANOVA found that task type had a statistically significant effect on EFL vocabulary knowledge, $F(3, 55.32) = 12.37, p < 0.001$. Table 3 reports the results of each linear contrast.

Table 3

Tests of task differences in EFL vocabulary knowledge

	$\hat{\psi}$	<i>p</i>	<i>p.crit</i>	<i>d_r</i>
Matching-Choice	-0.69	0.579	0.017	0.13
Matching-Definition	-8.87	< 0.001	0.013	1.27
Matching-Combining	-0.52	0.697	0.025	0.10
Choice-Definition	-8.18	< 0.001	0.010	1.13
Choice-Combining	0.17	0.887	0.050	0.03
Definition-Combining	8.35	< 0.001	0.008	1.14

$\hat{\psi}$ indicates a trimmed mean difference; *p.crit* refers to the critical value for a test of significance; + indicates significance at the specified critical *p* value.

As Table 3 indicates, all the differences between the definition task and any of the other tasks reach statistical significance, with very large effect sizes. All the differences among the matching, choice and combining tasks fail to reach statistical significance, with very small effect sizes.

Effects of Task Type on EFL Vocabulary Knowledge at Different Levels of EFL Proficiency

The previous section did not examine whether the task type effects were independent of EFL proficiency. In this section, a robust ANCOVA was conducted to examine the task type effects at each of the three EFL proficiency design points. These design points were 61.75, 69.25 and 74.71, representing the low, intermediate and high EFL proficiency level, respectively. Table 4 presents the results of the tests of task differences in EFL vocabulary knowledge at each design point.

Table 4

Tests of Task Differences in EFL Vocabulary Knowledge at Each EFL Proficiency Design Point

	Design point	<i>n</i> vs. <i>n</i> ₂	$\hat{\psi}$	<i>p</i>	<i>p.crit</i>	<i>d_r</i>
Matching-Choice	61.75	23 vs.19	-2.85	0.197	0.025	0.41
	69.25	30 vs.33	-0.70	0.661	0.025	0.12
	74.71	29 vs.30	0.90	0.497	0.017	0.19
Matching-Definition	61.75	23 vs.10	-9.63	0.000	0.010	1.55
	69.25	30 vs.30	-8.28	< 0.001	0.01	1.36
	74.71	29 vs.22	-9.41	< 0.001	0.013	1.41
Matching-Combining	61.75	23 vs. 21	0.46	0.792	0.050	0.08
	69.25	30 vs.27	0.09	0.918	0.050	0.02
	74.71	29 vs.22	0.88	0.550	0.025	0.17

	61.75	19 vs. 10	-6.78	0.029	0.013	0.82
Choice-Definition	69.25	33 vs.30	-7.58	0.0005	0.013	1.07
	74.71	30 vs.22	-10.31	< 0.001	0.010	1.55
	61.75	19 vs. 21	3.31	0.138	0.017	0.46
Choice-Combining	69.25	33 vs.27	0.79	0.618	0.017	0.12
	74.71	30 vs.22	-0.02	0.985	0.050	0.01
	61.75	10 vs. 21	10.09	0.000	0.008	1.55
Definition-Combining	69.25	30 vs.27	8.37	< 0.001	0.008	1.25
<i>g</i>	74.71	22 vs.22	10.29	< 0.001	0.008	1.44

ψ^{\wedge} indicates a trimmed mean difference; *p.crit* refers to the critical value for a test of significance; + indicates significance at the specified critical *p* value.

As Table 4 indicates, at each EFL proficiency level, the definition task outperforms all the other tasks significantly, with a large or very large effect size. No significant difference was found among the matching, choice and combining tasks at each EFL proficiency level. Nevertheless, at the low EFL proficiency level (design point = 61.75), given the observed low to medium-sized and medium-sized effect sizes, the small advantage of the choice task over the matching and combining tasks should be taken note of.

Discussion

This study investigated the task type effects on EFL learners' vocabulary knowledge, followed by examination of whether such effects were independent of EFL proficiency. Of the four types of task investigated, the definition task was found to outperform all the other tasks in promoting EFL vocabulary knowledge, and the matching, choice and combining tasks performed equally well. When EFL proficiency was considered, the definition task was also found to outperform all the other tasks at each level of EFL proficiency, and the matching and combining tasks showed no difference, irrespective of EFL proficiency. The choice task had an advantage over the matching and combining tasks at the low level of EFL proficiency, but not at the intermediate or high level of EFL proficiency. The task type effects are largely independent of EFL proficiency, a finding similar to Kim (2008). It seems that, once low-proficiency EFL learners have reached a level of proficiency sufficient to allow them to complete common vocabulary-focused tasks, it is possible for them to benefit from the tasks as equally or nearly equally as high-proficiency EFL learners. It should be noted, however, that the EFL learners in this study were at the same grade level, and their differences in EFL proficiency might not be as large as they appeared. Thus, it would be of great interest to examine whether the present findings would still hold across a wider range of proficiency levels.

With regard to the ILH, only partial support was provided, since all the four tasks were not equally effective in promoting vocabulary knowledge. The task type effects allow for explanations other than the ILH. In what follows, the effects are explained mainly in terms of differential processing and division of attention.

Differential Processing

The word exposure frequency effect, evidenced by some studies (e.g., Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2011) may partly account for the present findings. Take for example the comparison between the definition and combining tasks. The definition task increased the frequency at which each target word was encountered and evaluated, thus reinforcing the form-meaning connection in the learner's mental lexicon. Such frequent encounters of the target words could not be expected to occur in the combining task, where target words were given individually, and no comparisons between them were needed. The word exposure frequency effect might also account for the small advantage of the choice task over the matching or combining task when the learners were compared at the low EFL proficiency level. The choice task, where the target word appeared in each of the four options, might offer the learners multiple exposures to the target word. Compared to the learners at the intermediate or high EFL proficiency level, those at the low EFL proficiency level were more likely to repeatedly compare the four options concerning the target word in order to make a

correct choice. Consequently, multiple exposures to the target word might have increased the learners' chances of knowing the word.

Perhaps more importantly, the superiority of the definition task over the other tasks was due to more retrieval or mental effort. The learners doing the definition task had to infer a conceptually familiar L1 word equivalent corresponding to each definition/description, and then return to the reading exercises to find the target word whose actual L1 equivalent was the same as or similar to the inferred one. In order to ensure the form-meaning correspondence, the learners might have compared and evaluated the target words and the definitions/descriptions repeatedly. The forced output facilitated the definition task learners' access to the target word forms. The mental effort demanded by inferring and forced production might have greatly strengthened the form-meaning connection. This also explains Wesche and Paribakht's (2000) finding that interpretation of form-meaning relationships in the definition task induced more mental effort than recognition of form-meaning relationships in the matching task.

In the final analysis, differential processing could probably account for task effectiveness for vocabulary learning. Unlike the other tasks, the definition task could induce both elaborate structural (i.e., orthographic) and elaborate semantic processing of the target words. Although the matching task provided the same rich contextual cues as did the definition task, the learners on the matching task might have found no need to infer the target words, which had already been given. Even though the matching task learners might compare or evaluate the different target words and their corresponding definitions/descriptions, such comparisons or evaluations did not warrant the same semantic elaboration as did the definition task. A similar case is made for the choice and combining tasks. Like the matching task, these two tasks did not require effortful retrieval but recognition of the target words. The comparison between the definition and the other tasks reveals the importance of both structural and semantic processing in vocabulary learning.

Division of Attention

One may still wonder why the combining task was not more effective for vocabulary learning than the choice task. Intuitively, it appears reasonable to anticipate that the combining task would perform better than the choice task, not only because the former took longer than the latter, but also because the former was a recognition task, whereas the latter was a production one. More time on task does not necessarily lead to better retention of the target words, however. For instance, Craik and Tulving (1975) compared reaction time among three levels of processing, i.e., questions concerning type-script (structural level), rhyme questions (phonemic level), and sentence questions (semantic level), finding that slow responses were recognized little better than fast responses at each level of processing. In the same vein, Hill and Laufer (2003) found that task effectiveness could be attributed to task type rather than time on task. In the current study, the choice and combining tasks proved to be more complex than the matching task or perhaps even the definition task, because the learners doing these tasks, especially those doing the combining task, had to understand the meanings of all word strings and analyze the syntactic relations among them so as to rearrange the word strings into a proper sentence. However, like time on task, task complexity does not necessarily contribute to better word retention, either. For example, Joe (1998) found that an experimental task, where the adult L2 learners received explicit instruction on generative tactics and retold the passage without the aid of the text, did not perform better in word retention than a comparison task, where the learners did not receive explicit instruction but had the text available to them while retelling, although the experimental task was expected to outperform the comparison task. The increased task demands and higher learning burden imposed on the experimental group may have led to their failure to outscore the comparison group (Joe, 1998, p. 373).

In this study, although processing the target word was intended to be the primary activity, and unscrambling word strings into a proper sentence the secondary activity, the choice and combining tasks, especially the combining task, were so demanding (e.g., in terms of syntactic knowledge) that the learners might have diverted their attention to the secondary activity, resulting in shallower encoding of the target word processed in the primary activity. Still possibly,

various requirements dictated by the combining task depleted the learners' attentional resources, and this in turn led to faint memories of the new words which could possibly have been processed elaborately. This may explain why the combining task had no advantage over the matching or choice task. It should be noted that, though, the participants in this study were not advanced EFL learners, and their syntactic competence was not developed fully. Equipped with more syntactic competence, EFL learners doing the combining task would have been less likely to divert much attention to those requirements not directly related to the target words, and thus might have memorized the target words better, since the sentence context could have consolidated the form-meaning connection, or even the word usage, a claim that remains to be verified.

Conclusion

This study examined EFL learners' initial vocabulary learning through reading sentences and performing tasks related to the target words. Four tasks, involving two input (matching and choice) and two output tasks (definition and combining), were compared in EFL vocabulary knowledge acquisition. The definition task outperformed all the other tasks across EFL proficiency levels. The matching and combining tasks performed equally well, regardless of EFL proficiency. The choice task gained a small advantage over the matching and combining tasks when the learners were at the low level of EFL proficiency, but this advantage disappeared when the learners arrived at the intermediate or high EFL proficiency level.

This study contributes to a better understanding of what characteristics of a word-focused task determine its effectiveness for word retention. Differential processing may well account for the superiority of the definition task over the other tasks in retention of new words, suggesting the importance of both structural and semantic processing. Thus, in order for the ILH to better explain or predict task effectiveness, the notion of evaluation may need to be extended to cover both structural and semantic elaboration instead of the latter only. Division of attention was another probably important factor affecting word retention, since the task demands could direct or divert the learners' attention to the new words. Therefore, although all vocabulary learning tasks can be labeled as "word-focused", it does not follow that they would induce the same amount of learners' attention to new words. Divided attention, which depends on the task requirements and complexity, would probably reduce word retention even if the new words were once processed more deeply or elaborately.

These findings have potential implications for EFL vocabulary instruction. EFL teachers should keep in mind that neither more task time nor a more complex task is necessarily beneficial for vocabulary learning, since word retention depends much more on what kind of processing the task elicits and how much of the learners' attention the task directs to the new words. They are advised to design tasks which can induce access to both word form and meaning to help the learners consolidate the form-meaning connection. EFL teachers should also be aware that, at the initial stages of EFL vocabulary learning, attention to the structural properties of a new word may be of vital importance, since EFL vocabulary acquisition "typically does not involve learning new concepts while learning new word forms", as is true of L1 vocabulary acquisition (Barcroft, 2002, p. 356). Specifically, in designing vocabulary learning tasks with identical involvement loads, EFL teachers are encouraged to employ tasks like the definition task more frequently than those like the matching task if the primary pedagogical goal is to improve vocabulary learning. When the learners are just beginning to encode new EFL words, tasks like the combining task is not recommended, for those tasks may exhaust learners' processing and attentional resources which could otherwise be employed to process new word forms. If improving EFL learners' syntactic ability is the main pedagogical goal, however, tasks like the choice or combining task might be in order, with acquisition of new words being a by-product. For learners with low EFL proficiency, the choice task seems preferable to the combining task in contributing to vocabulary learning.

Notes

- 1 The 20% trimmed mean is computed by first removing 20% of the smallest and largest sample values and then averaging what remains.

- 2 The 20% Winsorized variance is the variance of the 20% Winsorized values, which are derived by pulling the smallest 20% of the sample observations up to the smallest value not trimmed, and the largest 20% of the sample observations down to the largest value not trimmed.
- 3 A robust ANCOVA picks points (design points) on x (covariate), and then compares the 20% trimmed y (dependent variable) means for all independent groups, based on the y values for each group corresponding to the x values in the neighborhood of each design x point. A robust ANCOVA allows nonnormality, heteroscedasticity, and even curved regression lines. The Benjamini-Hochberg method was used to control the family-wise error rate at each design point. See Wilcox (2017) for more information.

Acknowledgments

This article was supported by Jiangsu Social Science Foundation (Grant No. 18YYB013). I appreciated the anonymous reviewers' insightful comments and suggestions on an earlier version of this manuscript.

References

- Bao, G. (2015). Task type effects on English as a Foreign Language learners' acquisition of receptive and productive vocabulary knowledge. *System, 53*, 84-95.
- Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning, 52*(2), 323-363.
- Bird, S. (2012). Expert knowledge, distinctiveness, and levels of processing in language learning. *Applied Psycholinguistics, 33*(4), 665-689.
- Brown, T. S., & Perry, F. L. Jr. (1991). A comparison of three learning strategies for ESL vocabulary acquisition. *TESOL Quarterly, 25*(4), 655-670.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.
- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294.
- Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research, 16*(2), 227-252.
- Folse, K. S. (2006). The effect of type of written exercise on L2 word retention. *TESOL Quarterly, 40*(2), 273-293.
- Golonka, E., Bowles, A., Silbert, N., Kramasz, D., Blake, C., & Buckwalter, T. (2015). The role of context and cognitive effort in vocabulary learning: A study of intermediate-level learners of Arabic. *The Modern Language Journal, 99*(1), 19-39.
- Huang, L.-L., & Lin, C.-C. (2014). Three approaches to glossing and their effects on vocabulary learning. *System, 44*, 127-136.
- Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL, 41*, 87-106.
- Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539-558.
- Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*(3), 357-377.
- Keating, G. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365-386.
- Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285-325.
- Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716.
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1-26.

- Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, 15(4), 391-411.
- Min, H. (2008). EFL vocabulary acquisition and retention: Reading plus vocabulary enhancement activities and narrow reading. *Language Learning*, 58(1), 73-115.
- Paribakht, T. S., & Wesche, M. B. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174-200). Cambridge: Cambridge University Press.
- Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics*, 33(1), 66-82.
- Sagarra, N., & Alba, M. (2006). The key is in the keyword: L2 vocabulary learning methods with beginning learners of Spanish. *The Modern Language Journal*, 90(2), 228-243.
- Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33-52.
- Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, 43(1), 55-77.
- Wesche, M. B., & Paribakt, T. S. (2000). Reading-based exercises in second language vocabulary learning: An introspective study. *The Modern Language Journal*, 84(2), 196-213.
- Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). San Diego, CA: Elsevier.

About the Author

Gui Bao, PhD, is a Professor of Applied Linguistics at Nanjing Tech University, Nanjing, China. His research interests cover second language research methodology, second language vocabulary acquisition and applied statistics for linguistics.