English Language, Literature & Culture
2022; 7(1): 19-29
http://www.sciencepublishinggroup.com/j/ellc
doi: 10.11648/j.ellc.20220701.14
ISSN: 2575-2367 (Print); ISSN: 2575-2413 (Online)

Effects of Sentence Output Tasks on EFL Vocabulary Learning from a Bayesian Perspective

Gui Bao

School of Foreign Languages and Literature, Nanjing Tech University, Nanjing, China

Email address:
boggy2008@126.com

To cite this article:
Gui Bao. Effects of Sentence Output Tasks on EFL Vocabulary Learning from a Bayesian Perspective. *English Language, Literature & Culture*. Vol. 7, No. 1, 2022, pp. 19-29. doi: 10.11648/j.ellc.20220701.14

Received: January 20, 2022; **Accepted**: February 7, 2022; **Published**: February 19, 2022

Abstract: This article builds on the growing line of inquiry into the relative effectiveness of different tasks for vocabulary learning. Specifically, the study compares the efficacy of several sentence output tasks in EFL vocabulary learning through reading. To this end, evidence was weighed for one hypothesis over the alternative from a Bayesian perspective rather than in light of the commonly used null hypothesis significance testing (NHST), which depends heavily on the *p* values for statistical conclusions. Ninety-one EFL learners were randomly assigned to one of three word-focused sentence output tasks (i.e., L2-L1 translating, paraphrasing and writing), and were subsequently tested on their initial learning and retention of newly-encountered EFL words. Both Bayes factor analysis and Bayesian parameter estimation were employed to find evidence for task effects. With respect to initial word learning, moderate evidence was in favor of no difference between translating and paraphrasing, whilst weak evidence in favor of no difference between translating and writing as well as between paraphrasing and writing. For pedagogical purposes, no good evidence was for or against task effects. Regarding word retention, moderate evidence supported no task difference, and all the tasks fared equally well pedagogically. The results partially support the involvement load hypothesis and are discussed in terms of task difficulty, context generation and semantic elaboration.

Keywords: Vocabulary Learning, Sentence Output, Involvement Load Hypothesis, Context, Bayesian Methods

1. Introduction

It is universally acknowledged that L2 (second language) reading with word-focused tasks (i.e., reading plus) is more facilitative of L2 vocabulary learning than L2 reading without such tasks (i.e., reading only) [2, 13, 27, 37]. Word-focused tasks through reading draw the learners' attention to new words, thus increasing the chances that they 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 ([13], p. 90).

L2 research on word-focused tasks has gained momentum, particularly since the inception of the levels of processing framework [6, 7] and more notably the involvement load hypothesis (hereafter ILH) [15, 23]. Both the levels of processing framework and the ILH underscore the crucial role of semantic elaboration in word learning. Semantic elaboration refers to any mental operation or evaluation of a word (or a vocabulary item) with regard to its meaning [4]. It entails, among others, mentally connecting a new word with ones already known, embedding the word in a meaningful context and/or associating the word with a mental image. Semantic elaboration can be strong or weak. When required to write a new sentence with a newly-encountered word, for instance, the learner will have to take account of the word meaning and usage and decide on how the word combines with other words to produce a meaningful and well-formed sentence, and thus strong elaboration on the word will necessarily be generated. Gap-filling, where the learner is required to choose a word from a word list that fits a given sentence best, entails less strong elaboration, because it requires a comparison of word meanings only, not stretching the learner's linguistic resources as far as sentence writing does. Following the levels of processing framework and the ILH, more elaboration on the word should put sentence writing at an advantage over gap-filling in facilitating vocabulary learning.

Regarding the task-related context, two predictions arise from the ILH [23]. One is that an original context, as demanded by word-focused writing, for instance, will

contribute more to L2 vocabulary learning than a given context, as demanded by gap-filling. The other is that word-focused writing like sentence writing and composition/passage writing will induce the same degree of semantic elaboration or evaluation, resulting in similar amounts of L2 vocabulary learning. Although studies of gap-filling, sentence writing and/or passage writing are numerous [11, 16, 17, 25, 32, 38], there is a paucity of research on comparing sentence translating and writing [1] as well as on comparing gap-filling and sentence paraphrasing (rewriting) [14]. Of special concern, little research to date has been conducted on the relative efficacy of translating, paraphrasing or writing in the sentence context. This line of research would advance L2 educators' understanding of how similar output tasks affect L2 vocabulary learning and whether the ILH still holds across these tasks. This study addresses how these sentence output tasks would affect EFL (English as a Foreign Language) initial word learning and retention.

2. Literature Review

The levels of processing framework and the ILH are frequently employed to explain the relative efficacy of tasks for vocabulary learning. As the levels of processing framework suggests, semantic processing (deep processing) will be more conducive to vocabulary learning than structural (e.g., orthographic or phonetic) processing (shallow processing), and more elaboration at the semantic or structural level will contribute more to vocabulary learning [6, 7]. In the field of L2 vocabulary research, The ILH constitutes an initial attempt to define notions like "depth of processing" and "degree of elaboration" unambiguously. The ILH ascribes 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 [23], "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 an L2 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 there is no attempt, or present (+S) if there is. "Evaluation" involves a decision on the word's form or meaning. "Evaluation" can be absent when the task is not relevant to the new words, or moderate when the task entails recognizing differences between words (as in a gap-filling task), or a decision on the meaning of a polysemous word in a given context. Strong "evaluation" involves a decision as to how a new word will combine with other words in an original L2 context. The absence of a component is marked as 0, moderate presence as 1, and strong presence as 2. The involvement indexes can be added to represent the degree of overall involvement. The ILH posits that, the greater the task-induced involvement loads, the more likely the word will be learned. It also suggests that tasks with identical involvement loads will be equally effective in enhancing L2 word learning.

An issue of interest is whether the originality of a task-related context will necessarily make a difference in L2 vocabulary learning. The ILH suggests that, when other conditions ("need" and "search") are held constant, an original context will benefit L2 vocabulary learning more than a given context. Gap-filling, for instance, induces moderate evaluation (+E), because "the use of words is evaluated in given contexts" [21]. L2-L1 sentence translating also induces moderate evaluation (+E), since the target word is "evaluated against the other words surrounding it" and a decision is made "after an evaluation of several translation alternatives" ([22], p. 712). In contrast, both sentence writing and passage writing (composition or summary) induce strong evaluation (++E) in the sense that these tasks entail the use of a target word in an original text. Similarly, sentence paraphrasing or rewording also induces strong evaluation (++E), since in sentence paraphrasing as in L1-L2 translating, the entire L2 context will be created by the learner ([22], p. 712), although it should be added that sentence paraphrasing and sentence writing may elicit different degrees of contextual originality or novelty. The superiority of sentence/passage writing over gap-filling was found in some studies [15, 17, 36, 38] but not in others [11, 25]. Contrary to the direction predicted by the ILH, gap-filling was found superior to sentence paraphrasing (rewriting) in [14]. Sentence writing fared as well as sentence translating in [1] as well as passage writing in [17] but not in [38] or in [12]. Rassaei (2017) [32] found that several variants of word-focused composition (i.e., text summarization, generating questions out of the texts and answering them, and making predictions about what is to occur in the texts) did not always perform equally well. It seems that how much elaborative processing relates to the learner-generated context is more complicated than assumed by the ILH.

A methodological issue regarding testing the ILH is the use of the conventional or traditional null hypothesis significance testing (NHST), a frequentist perspective. The interested reader is referred to [3, 29, 30] for more about Bayesian and frequentist methods. Briefly, the p value delivered by NHST can provide direct evidence for an alternative hypothesis (H_1) (e.g., there is a task effect), but not for a null hypothesis (H_0) (e.g., there is no task effect). The interested reader is referred to [9] for more about NHST. Essentially, NHST is conditional upon the presupposition that H_0 is true. A significant result (usually $p < 0.05$) can be considered as support for H_1. A non-significant result (usually $p > 0.05$), however, cannot be regarded as support for H_0. In this regard, the p value, no matter how large it is, cannot distinguish evidence for H_0 from no evidence at all. With respect to testing the ILH, NHST would be inappropriate if one intended to find evidence in favor of the null hypothesis that tasks with identical involvement loads would contribute equally to L2 word learning.

Taken together, this literature review suggests that uncertainty remains as to the relative effectiveness of sentence output tasks. There is a clear need to explore how sentence output tasks will affect EFL learners' vocabulary learning and retention.

To that end, the present study addressed the following questions:

1. How do sentence output tasks (i.e., L2-L1 sentence translating, sentence paraphrasing and sentence writing) affect EFL learners' initial learning of EFL word meaning?
2. How do sentence output tasks (i.e., L2-L1 sentence translating, sentence paraphrasing and sentence writing) affect EFL learners' retention of EFL word meaning?

According to the ILH, there would be no difference in initial EFL word learning or in EFL word retention between sentence paraphrasing and sentence writing, but there would be a difference when L2-L1 sentence translating was compared with both sentence paraphrasing and sentence writing. Since "proving" the no-difference hypothesis (i.e. H_0) was intended as one research purpose, this study tested the null and alternative hypotheses from a Bayesian perspective rather than in light of NHST. Bayesian hypothesis testing does not assign a special status to the null hypothesis, but rather weighs evidence for the null hypothesis just as for the alternative one. Bayesian hypothesis testing can demonstrate that "the null hypothesis is more credible than the alternative hypothesis, which NHST can never do" ([20], p. 196).

3. Method

3.1. Research Design

This study used a between-subjects design to investigate how sentence output tasks would affect EFL learners' acquisition of vocabulary knowledge. Sentence output tasks included L2-L1 translating, paraphrasing and writing. These tasks induced "need," since there was a need to use target words, but no "search," since the word glosses were given. Following the ILH, L2-L1 sentence translating induced a moderate evaluation (+E), whilst sentence paraphrasing and sentence writing each a strong evaluation (++E) (See Section 2).

EFL vocabulary knowledge would be measured twice, one for initial word learning and the other for word retention. Both measures focused on each target word's L1 (Chinese) and L2 (English) meaning. Both L1 and L2 meaning knowledge would be measured because they were included in the word glosses.

During-task performance was also measured to examine whether task difficulty might be confounded with task-induced involvement loads in contributing to L2 word learning.

3.2. Participants

Participants were a total of 91 first-year English learners from three parallel intact English classes at a university in China. They were made up of 44 males and 47 females, aged from 17 to 20 years old (M = 18.23, SD = 0.62). All of them had received at least six years of formal EFL instruction before entering university. These participants were randomly assigned to one of the three output tasks, i.e., translating (n = 30), paraphrasing (n = 30) and writing (n = 31). One week prior to the experiment, they were given a word recognition test of English vocabulary, which was adapted from the Vocabulary Levels Test (Version 2) [35]. On the test were 150 words, the response to each of which was scored either 1 or 0 point. The mean scores for all groups ranged from 90.97 to 88.39 out of a full score of 150 points. A Bayes factor analysis (see Section 3.7 for more information) found moderate evidence for no mean difference in EFL vocabulary size among the task groups (BF_{01} = 4.91). This partly indicates that the random assignment was reasonably good.

3.3. Materials

3.3.1. Reading Passage and Target Words

The paper-based reading passage for this study was adapted from "My Devilish Older Sister" in [28]. Major changes involved removing the title and replacing the blanks with target words in boldface. The adapted passage was 309 words long at a Flesch-Kincaid Grade Level of 6. The passage was accompanied by five comprehension questions, none of which focused upon the target words. The main purpose of these questions was to stimulate the participants to read the passage before proceeding to the output tasks.

Given the time limit on task-related vocabulary learning through passage reading, a range of approximately 10-20 target words was deemed appropriate, for instance, 10 words in [17], 14 words in [14, 32]. To this end, an experienced university EFL teacher was invited to read the adapted passage, and choose 20 English words that most participants were unlikely to know. These words, together with 11 distracters from the reading passage were tested with 18 non-participants from a parallel English class two weeks before the experiment. Consequently, 14 target words, whose correct L1 (Chinese) meanings they all failed to produce, were selected for this study. Among the target words were an equal number of verbs (*relinquish*; *forestall*; *insinuate*; *permeate*; *inundate*; *interrogate*; *torment*) and adjectives (*omnipotent*; *grotesque*; *obsequious*; *opportune*; *diabolic*; *insidious*; *contrite*).

The test results also led to some minor changes in the passage to facilitate the participants' understanding, such as easy substitutes for difficult words as well as in-text Chinese equivalents of a couple of difficult English words. The pretest of the target words, though, was not given to the participants in the experiment not only because the test might sensitize them to these words so that bias might be introduced to the posttest results but also because the non-participants and the participants were comparable in English proficiency so that their knowledge of the target words could not differ greatly from each other.

3.3.2. Sentence Output Tasks

Ten sentences wherein one or more target words occurred were extracted from the reading passage. The target word(s) in each sentence was (were) in boldface across all EFL vocabulary learning tasks. The translating task required the participants to translate each sentence from English to Chinese, while the paraphrasing task asked the participants to explain each sentence in their own English words, disallowing repetition of the target word(s). The writing task asked the participants to learn each sentence carefully before writing a new English sentence, which included a target word, showing its meaning.

The target words in the passage were glossed bilingually. The glosses consisted of each target word's Chinese meaning, English meaning, pronunciation in IPA, and part of speech, as is widely used in the Chinese context. They were placed in the right-hand margins of the passage.

3.3.3. Word Knowledge Posttests

In this study, the participants' initial learning of the target words was measured on an immediate test of the Chinese and English meanings of each of 14 target words. During the test, the participants were asked to write down the English meaning of each target word and its Chinese equivalent. The participants' word retention was measured on a delayed test, which was conducted one week after the immediate posttest. The target words on the delayed posttest were randomized in a different order from those on the immediate posttest to mitigate the memory effects.

3.4. Piloting

The passage reading, output tasks and vocabulary knowledge posttest were pilot-tested with the non-participants one week after the word meaning production test was given. The purpose of the pilot test was twofold. First, the test resulted in minor modifications of the instructions based on the students' feedback. For example, some students said that examples would help them understand the English instructions better, so an example with "disparity," an unfamiliar non-target word whose Chinese meaning was given in the reading passage, was presented on each posttest to assist the participants with understanding. Second, the maximum time limits for the passage reading plus each task and for the posttests were determined. The maximum amount of time for the passage reading plus each task was 25 min, with 7 and 18 min for the passage reading and each task, respectively. A total of 8 min was set on either of the immediate and delayed posttests.

3.5. Procedures

One week before the experiment, the researcher explained instructions and general research purposes to three research assistants, and trained them to hand out and collect the materials. The participants were not told that they were participating in a vocabulary learning experiment. They performed the output tasks during normal class time. During the experiment, each participant was asked to do the same reading comprehension exercises before proceeding to one of the three output tasks as per the instructions. While performing these tasks, all the participants had access to the word glosses in the reading passage.

Upon collection of all the worksheets, the research assistants administered an unannounced word knowledge posttest immediately. One week later, all the participants were given another unannounced word knowledge posttest. The teachers informed the researcher that the participants did not practice the target words in class between the immediate and delayed posttests.

A manipulation check of the worksheets found that the overwhelming majority of participants (74 out of 91) worked on at least 10 out of the 14 target words during the tasks, so the implementation fidelity was guaranteed.

3.6. Scoring

The L1 and L2 meaning tests were both scored trichotomously. Specifically, the L1/L2 meaning of a target word was awarded 1 point if it was closely aligned with the correct meaning of the word, 0.5 point if it was partly correct, or 0 point if it was totally incorrect or there was no response. For example, the L1 meaning of "insidious" was awarded 1 point if the word was translated into "暗中有害的," and the L2 meaning was awarded 0.5 point when the word was explained as "harm," since "harm" partly conveyed the meaning of "insidious."

Each participant's during-task performance was also scored trichotomously. For each task, 0.5 point was awarded if a target word was translated properly, rephrased properly, or used in a meaningful new context. Another 0.5 point was awarded if the translation reproduced the original meaning and read naturally, the paraphrase involved no serious grammatical error, or the target word was used grammatically. However, 0 point was given in the presence of mistranslation, misinterpretation, misuse, or omission of the word. Thus, each target word could receive a score of 0, 0.5, or 1. No consideration was given to minor misspellings and grammatical errors unrelated to the target word.

Two EFL researchers were trained to score the participants' responses independently. For the L1 and L2 meaning tests, the inter-rater agreement was 100%, as all disagreements were resolved by consensus. For during-task performance, the inter-rater reliability was high (Cronbach's $\alpha = 0.97$), so the scores were averaged for each participant.

3.7. Data Analysis

The usual statistical method to address each research question in this study is a conventional one-way analysis of variance (ANOVA), a frequentist approach to amassing evidence to reject the null hypothesis (i.e., all group means are equal) using a preset alpha level (e.g., $\alpha = 0.05$). One problem with this approach is that it tells us nothing about the probability that the null hypothesis adequately represents the data or the probability that the alternative hypothesis (H_1, i.e.,

the means are not equal) does a better job. Another problem with this approach is that, when the *p* value from an ANOVA is greater than the alpha level, we fail to reject the null hypothesis, but we cannot accept the null hypothesis blindly. A third problem with this approach is that a significant *p* value can readily let us slip into the fallacy of equating statistical significance with practical (or pedagogical) significance, although the *p* value in itself indicates nothing about a meaningful mean difference or effect size. To overcome the limitations of a conventional ANOVA, this study employed a Bayesian approach.

In this study, the research questions regarding task contrasts in EFL word knowledge means were each addressed in two complementary ways so as to fully understand the relative efficacy of output tasks. A Bayes factor analysis was conducted to find evidence in favor of the null hypothesis over the alternative or vice versa. To further ascertain whether the mean differences across the output tasks were practically or pedagogically meaningful, this study did Bayesian parameter (i.e., population mean difference) estimation with a 95% highest density interval (HDI) plus a region of practical equivalence (ROPE).

The Bayes factor quantifies the strength of evidence in favor of one hypothesis over the alternative one. BF_{01} indicates the strength of evidence in favor of H_0 (the null hypothesis) over H_1 (the alternative hypothesis). BF_{10}, inversely related to BF_{01}, indicates the strength of evidence in favor of H_1 over H_0. Put differently, the Bayes factor indicates which of the two rival hypotheses is more likely or plausible given the observed data. For instance, $BF_{01} = 2$ demonstrates that H_0 is twice as likely as H_1. Unlike the *p* value from a conventional ANOVA, which quantifies evidence for H_1 but not for H_0, a *BF* provides a symmetrical measure of evidence for H_1 versus H_0; that is, there can be evidence for H_0 just as much as for H_1, or insufficient evidence either way. This study used the following guidelines for interpreting a BF_{01}: 1, no evidence; 1-3, anecdotal or weak evidence; 3-10, substantial or moderate evidence; 10-30, strong evidence; 30-100, very strong evidence; > 100, decisive evidence [30]. To illustrate, $BF_{01} = 25$ suggests strong evidence for H_0 versus H_1.

For practical or pedagogical purposes, statistical decisions based on Bayesian parameter estimation involved a comparison of a 95% HDI with a ROPE. A 95% HDI spans 95% most probable mean differences in a task contrast, including the median of mean differences. It has an intuitive interpretation: there is 95% probability that the mean difference in a task contrast falls within the range defined by the HDI. The mean differences within a 95% HDI are more credible than those outside it. Likewise, the mean differences in the middle of the HDI tend to be more credible than those at the limits of the HDI.

A ROPE covers all population mean differences that may be considered too small to be meaningfully different from zero. In this study, we were interested more in pedagogically meaningful mean differences than in exactly no differences in task contrasts. The ROPE was defined as an interval between -2 and 2. Here, -2 and 2 refer to a difference of 2 points out of a total word knowledge raw score of 28 points; 2 points could suggest successful acquisition of a new word meaning in both L1 and L2. For initial EFL word learning, a mean difference of 2 points is roughly equivalent to Cohen's *d* of 0.4, a value slightly lower than a medium effect size of Cohen's $d = 0.5$. For EFL word retention, a mean difference of 2 points is roughly equivalent to Cohen's *d* of 0.5. Using a ROPE allows us to get a probability for a specific hypothesis, such as "The mean difference between sentence translating and sentence paraphrasing is less than 2 or larger than 2." To accept or reject the null hypothesis of no mean difference, we followed the HDI+ROPE decision rule [19]. Specifically, if the 95% HDI fell completely within the ROPE, the null hypothesis would be accepted for practical purposes. If the 95% HDI fell completely outside the ROPE, the null hypothesis would be rejected. Otherwise, the decision would be withheld.

This study computed Bayes factors via the program "ttestBF" from the R package "BayesFactor" developed by [26], and derived the 95% HDIs and ROPEs of the estimates of population mean differences across the tasks by running the program "Jags-Ymet-Xnom2fac-MrobustHet" developed by [18]. The posterior distributions of population mean differences were generated by the Markov chain Monte Carlo (MCMC) methods (a class of modern sampling techniques) using the "runjags" package [8]. All the data analyses were conducted by R 3.6.1 [33]. The Gelman-Rubin statistic was used to assess convergence of the MCMC chains.

4. Results

This section presents some descriptive statistics for task-related initial learning and retention of the target words' meanings, followed by a Bayes factor analysis and a Bayesian parameter estimation of the mean differences in each task contrast.

4.1. Initial Word Learning and Retention Across Output Tasks

The participants on all three output tasks (i.e., translating, paraphrasing and writing) were tested on the target words' meaning knowledge twice to elicit data for initial word learning and word retention, respectively. Figure 1 displays two sets of boxplots to compare the tasks in contributing to the target words' meaning knowledge on the immediate and delayed tests, respectively. Included in the figure are the sample size (*n*), mean (*M*) and standard deviation (*SD*) for each task.

In Figure 1, a rectangular box is drawn for each task group, extending from the lower quartile to the upper quartile, with the median shown by a thick line. Whiskers extend from both ends of the box to the greatest and smallest values. The values outside the whiskers, if any, are denoted with empty circles as outliers.

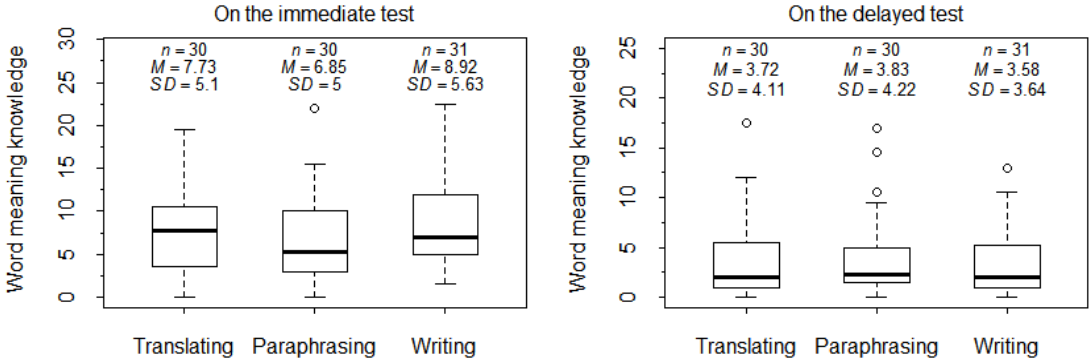

Figure 1. *Boxplot comparisons of the output tasks across time.*

As shown in the left panel, regarding initial word learning, the writing group has a slightly larger mean than the translating group, which in turn has a slightly larger mean than the paraphrasing group. The median lines, however, indicate that the translating group has a slightly larger median than the writing group (7.75 vs. 7 points), where the data distribution is positively skewed. The paraphrasing group has the lowest median (5.25 points), with its data distribution positively skewed with an outlier (a score of 22 points).

The right panel of Figure 1 compares the three tasks with respect to word retention. Considerable attrition of word meaning knowledge regardless of task type suggests that elaborative processing during the first encounter of new words is unlikely to facilitate long-term retention. To illustrate, the word meaning knowledge gains in the translating group, decrease from a mean of 7.73 points on the immediate test to a mean of 3.72 points on the delayed test. Consequently, all the task groups have virtually identical small means and medians of EFL word retention scores (the median for each task from the left to the right: 2, 2.25 and 2 points). In each group, one or more outliers positively skew the data distribution.

To sum up, the three task groups present small mean and median differences on both the immediate and delayed tests, especially on the latter.

4.2. Effects of Task Type on EFL Learners' Initial Word Learning

This section reports on the results from a Bayes factor analysis of the task effects on EFL learners' initial word learning, followed by an estimation of the posterior mean differences in each task contrast.

A Bayes factor analysis was conducted to test the relative task effectiveness in initial word learning. Moderate evidence was in favor of the null hypothesis of no mean difference between translating and paraphrasing (BF_{01} = 3.14), whereas weak evidence in favor of the null hypothesis of no mean difference between translating and writing (BF_{01} = 2.81) and between paraphrasing and writing (BF_{01} = 1.47).

Bayesian parameters were estimated to gauge the task effects on EFL learners' initial word learning. All the Gelman-Rubin statistics were less than 1.1, suggesting that the MCMC chains mixed well and converged to the desired posterior distribution. The posterior mean differences in each task contrast are displayed in Figure 2.

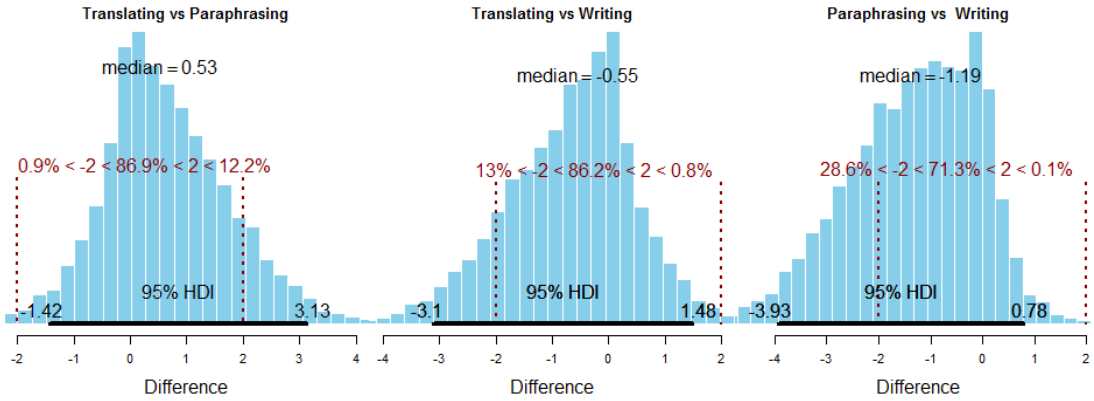

Figure 2. *Mean differences in initial EFL word learning.*

As shown in each panel of Figure 2, a horizontal axis indicates mean differences in word meaning knowledge-scale units. The 95% HDI is marked by a horizontal bar while the ROPE limits by two vertical lines at ±2 units in the word meaning knowledge-scale units. Also shown is the median of the posterior distribution of mean differences in a task contrast.

The leftmost panel demonstrates that the typical mean

difference between translating and paraphrasing is very small (a median of 0.53). As the 95% HDI indicates, there is a 95% probability that the mean differences extend from -1.42 to 3.13, with a mean difference of zero included in between. The ROPE, which covers 86.9% of all the mean differences, does not contain all the plausible mean differences within the 95% HDI. This suggests that the conclusion is equivocal. That is, in the sense of practical equivalence, the null hypothesis of no practical task difference cannot be accepted or rejected with good evidence. Likewise, the middle panel demonstrates that the typical mean difference between translating and writing is very small (a median of -0.55). The ROPE, which covers 86.2% of all the mean differences, does not contain all the plausible mean differences within the 95% HDI, suggesting inconclusiveness, too. As shown in the rightmost panel, writing tends to fare better than paraphrasing (a median of -1.19), since a majority of mean differences in the 95% HDI are less than zero. Nevertheless, the 95% HDI does not fall completely outside the ROPE, suggesting inconclusive evidence for the superiority of writing over paraphrasing for practical purposes.

4.3. Effects of Task Type on EFL Learners' Word Retention

This section presents the results from a Bayes factor analysis of the task effects on EFL learners' word retention, followed by an estimation of the posterior mean differences in each task contrast.

A Bayes factor analysis was conducted in the same way as in Section 4.2. Moderate evidence was found in favor of the null hypothesis of no mean difference in each task contrast (translating and paraphrasing: BF_{01} = 3.79; translating and writing: BF_{01} = 3.81; paraphrasing and writing: BF_{01} = 3.74).

As in Section 4.2, Bayesian parameters were estimated to gauge the task effects on EFL learners' word retention. The MCMC chains were found to mix so well, since all the Gelman-Rubin statistics were less than 1.1. The posterior mean differences in each task contrast are shown in Figure 3.

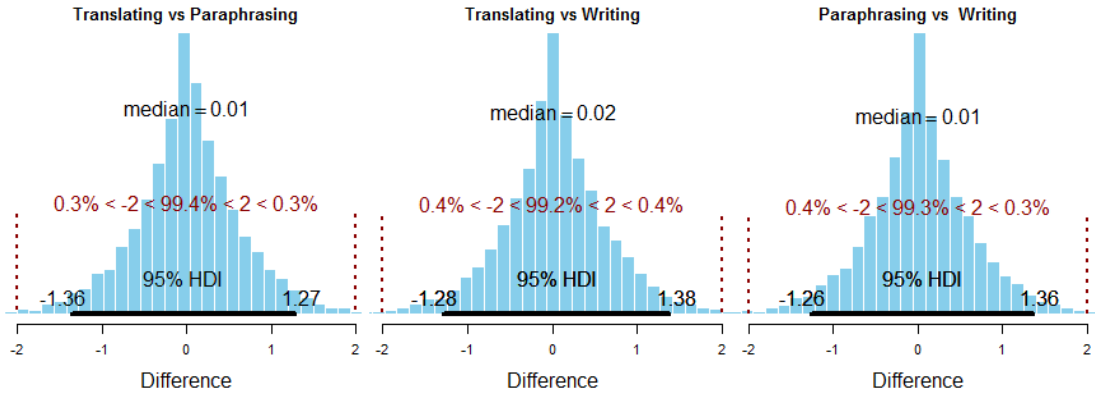

Figure 3. *Mean differences in EFL word retention.*

As shown in Figure 3, all the medians of mean differences in three contrasts are virtually zero; that is, each entire distribution of mean differences centers almost on zero. All the three HDIs span similar ranges with zero roughly in the middle. Moreover, all the three HDIs fall within a ROPE of ±2. This suggests that all the 95 % most credible values of the estimated mean differences are practically equivalent to zero, so the null hypothesis of no practical mean difference is accepted. To illustrate, the leftmost panel indicates that the ROPE, which covers 99.4% of all the mean differences between translating and paraphrasing, includes 95% most credible mean differences. To conclude, translating, paraphrasing and writing fare equally well or poorly in retention of the word meanings for pedagogical purposes.

5. Discussion

This study employed a between-subjects experimental design to answer the two questions regarding relative task effectiveness in EFL vocabulary learning. Three sentence-level output tasks (i.e., translating, paraphrasing and writing) were compared in improving initial word learning and word retention.

The first question asked whether there would be task effects on EFL initial word learning. The null hypothesis of no mean difference between translating and paraphrasing was accepted with moderate evidence. The other two null hypotheses regarding the writing-related contrasts were supported with weak evidence. For practical or pedagogical purposes, no good evidence was in favor of the superiority of one task over another.

The second question asked whether there would be task effects on EFL word retention. The null hypothesis of no mean difference in each task contrast was accepted with moderate evidence. Moreover, evidence supported no task effects for pedagogical purposes.

5.1. Task Difficulty, Contextual Originality and EFL Word Learning

The present findings resonate partly with [1], in which sentence translating from L2 to L1 and sentence writing were found to perform equally well in improving initial EFL word learning. Nation and Webb (2011) [27] claimed that sentence paraphrasing (rewording sentences, as they called) would lead to a less high degree of generation than sentence writing (i.e.,

sentence production) because "the degree of generation is constrained by the original sentence" (p. 9). Even though there may be some truth in their argument, this study does not lend support to their claim. This study did not lend full support to the ILH [23], either.

According to the ILH, translating would be less effective than both paraphrasing and writing, since translating was assumed to induce lower involvement loads than paraphrasing and writing, both of which were assumed to induce identical involvement loads. Based on the present findings about initial EFL word learning, the ILH was rejected in testing the null hypothesis of no mean difference regarding translating and paraphrasing, but left undecided in testing the other null hypotheses. For word retention, whether the ILH holds depends on which task contrast is considered. Specifically, the ILH was tenable when paraphrasing and writing were compared, but not when translating was compared with paraphrasing or writing.

Overall, the present findings seem to draw a vague picture of the comparative efficacy of sentence-level tasks in initial word learning. This is mainly because no conclusive evidence in this study was found for task effects except for the contrast between translating and paraphrasing. Given good evidence regarding word retention in this study, a clear picture is drawn; that is, all the three tasks fared equally well.

In an attempt to explore the possible causes of relative task efficacy, we examined task difficulty and its relationship with EFL word learning. A Bayes factor analysis found decisive evidence against the null hypothesis of no mean difference in during-task sentence production between translating and paraphrasing (for translating: M = 13.21, SD = 0.92; for paraphrasing: M = 10.93, SD = 2.12; BF_{10} = 11207.11) and between paraphrasing and writing (for writing: M = 12.90, SD = 1.12; BF_{10} = 784.33), but weak evidence in favor of the null hypothesis between translating and writing (BF_{01} = 2.18). In other words, paraphrasing turned out to be more difficult to the learners than both translating and writing, both of which were roughly at the same difficulty level.

As is not expected, task difficulty did not positively or negatively influence EFL word learning. A Bayesian correlation analysis was conducted to test the relationship between during-task performance and immediate EFL word learning. A positive correlation was supported with weak evidence on the translating task (BF_{10} = 2.40), whilst no positive correlation with weak evidence on both the paraphrasing and writing tasks (paraphrasing: BF_{01} = 2.42; writing: BF_{01} = 1.85). Similarly, a positive correlation was supported with moderate evidence between during-task performance and EFL word retention on the translating task (BF_{10} = 9.72), but no positive correlation with weak evidence on both the paraphrasing and writing tasks (paraphrasing: BF_{01} = 1.86; writing: BF_{01} = 0.90). We speculate that task difficulty will increase the learner's cognitive load and deplete his or her attentional resources, which might not be directed to a newly-encountered word and thus improve its memory.

Does contextual originality facilitate EFL word learning in the direction predicted by the ILH? Laufer and Girsai (2008) [22] assigned higher involvement loads (strong evaluation, ++E) to sentence writing than to sentence translating (moderate evaluation, +E), believing that strong evaluation was related to a totally new context as induced by sentence writing rather than sentence translating. The counterevidence from this study indicates that how much original the context is may not be as important as previously thought. Rather than contextual originality, how self-generated context relates to semantic elaboration might contribute to EFL word learning, which is to be discussed next.

5.2. Self-generated Context, Semantic Elaboration and EFL Word Learning

The learner's strong evaluation of or elaboration on a new word is closely related to his/her self-generated context where the word is embedded. The learners doing sentence writing generated an entirely original context for each target word, those doing sentence paraphrasing a partly original context, and those doing L2-L1 translating a context highly restricted in originality. All these output tasks, though, involved the learners' active engagement in producing a context for task completion and word learning. The learners across the tasks had to consider the grammatical relations among words, combine word chunks and generate a proper sentence in either L1 or L2. This process of sentence generation entailed much mental efforts and strong evaluation or semantic elaboration on the part of the learners. If the target word fitted a self-generated sentence context well in the learners' minds, the learners could possibly form a unified percept of the whole sentence, the memory for the target word could probably represent an integrated meaningful pattern, and thereby the word meaning could be retained equally well across the tasks.

Regardless of task difficulty and contextual originality, the three tasks in this study might be assumed to induce an identical involvement index of three (+N, ++E). It is suspected that the actual writing of the target word should have put the learners on the writing task at an advantage over translating and paraphrasing in strengthening the link between the word form and meaning. Merely repeating the target word once in writing, though, required little mental efforts or processing on the part of the learners. More importantly, all the three tasks in this study shared the following typical features: incorporating a productive or generative use of a target word into the task, the same amount of time on task, one receptive retrieval or search for a target word's meaning from the reading passage in case of the learners' recall failure, and a probable use of the self-generated sentence context for word retention. All these common task features should outweigh small between-tasks differences like contextual originality, and thus led to similar word learning outcomes.

Both the strength and limitation of the ILH lie in its simplicity. Some ideas suggested by the hypothesis stand in need of modification if it remains to be valid across various task conditions, especially the notion of contextual originality. We posit that both the task-induced involvement loads and the self-generated context would play a role in L2 vocabulary learning collaboratively and individually; that is, both the

involvement effect and the context effect may relate to L2 vocabulary learning. Although related, involvement and context may contribute to L2 vocabulary learning in different ways. The self-generated context would provide contextual clues for the memory for and retrieval of a newly-encountered word. The richer the clues are, the more accessible the target word will be. The task-generated involvement entails the learning burden of the target word and the during-task attention and mental efforts directed to the word. In this study, all the tasks demanded generation of separate sentences with the target words and entailed similar semantic elaboration on them, thus providing similarly rich contextual clues for later recall.

The synergic effects of involvement and context may well account for other research findings [11, 17, 38]. In testing the prediction by the ILH that tasks with identical involvement loads would result in similar vocabulary acquisition, Kim (2008) [17], for instance, compared the efficacy of sentence writing (+N, ++E) versus composition writing (+N, ++E) in enhancing initial word learning and retention. The participants were two groups of ESL (English as a Second Language) learners at different English proficiency levels, and the learners at each proficiency level were randomly assigned to either sentence writing or composition writing. The conventional ANOVAs showed that there was no significant main effect for either task type or proficiency level and also no task type-by-proficiency level interaction on the immediate and delayed posttests ($p > 0.05$). Using the statistics reported in [17], we performed a meta-analysis to find evidence for the null hypothesis versus the alternative hypothesis. Weak evidence was in favor of the null hypothesis about initial word learning ($BF_{01} = 2.82$), and word retention ($BF_{01} = 1.66$), so the ILH was partly supported. Zou (2017) [38] also compared the relative effectiveness of sentence writing and composition writing in an experiment with intermediate EFL learners, but found the superiority of composition writing over sentence writing in both initial word learning and word retention ($p < 0.001$). Using the statistics reported in [38], we found decisive evidence ($BF_{10} = 147.36$) in favor of the alternative hypothesis regarding initial word learning and very strong evidence for ($BF_{10} = 66.90$) in favor of the alternative hypothesis regarding word retention, thus rejecting the ILH. In [17], the sentence writing group was asked to write an original sentence with each given target word, whereas the composition group to write a passage (i.e., a one- to three-paragraph essay) about a topic, using the given target words. Since no other requirements like text organization were made, the composition group might have tried to work out individual sentences with the target words, paying little attention to how the target words were logically connected to each other at the passage level, especially under time pressure. In this regard, the passage context, which would otherwise have offered richer contextual cues for L2 meaning retrievals, could have been reduced to disconnected sentence contexts, as in the sentence writing task. Accordingly, similar context effects, coupled with identical involvement loads, probably led sentence writing and composition writing to similar amounts of ESL vocabulary learning. Zou (2017) [38] made the same requirements for sentence writing as in [17], but made more demanding requirements for composition writing by explicitly asking the EFL participants to coherently connect the target words in a passage so that they had to conceive a unified context to relate each target word to at least one of the others. The participants who wrote a composition did generate coherent sentence contexts to connect target words semantically when no time limit was set on the writing task, as the interview and think-aloud data in that study attested to. Therefore, multiple contexts for a target word as dictated by the composition writing requirements offered richer contextual clues for retrieving the word meaning, putting composition writing at an advantage over sentence writing in vocabulary learning. The comparison between [17] and [38] revealed that task effectiveness was determined not only by task-induced involvement loads but also by other task-related features like the degree of context generation or contextual richness.

6. Conclusion

This study extends previous research by examining the relative efficacy of sentence output tasks (L2-L1 translating, paraphrasing and writing) in EFL learners' initial word learning and word retention. With respect to initial word learning, weak or moderate evidence was found for no difference among the three tasks, and no good evidence was in favor of the superiority of one task over another for pedagogical purposes. More importantly, with respect to word retention, moderate evidence was found for no task difference, and good evidence supported no task effects for pedagogical purposes.

These findings drive us to reconsider the "evaluation" component in the ILH. This study suggests that, compared to L2-L1 sentence translating and sentence paraphrasing, sentence writing could induce slightly more orthographic elaboration, but all the three tasks could elicit the same degree of semantic elaboration. In this sense, whether "evaluation" is moderate or strong depends more upon how the learner makes use of the self-generated context to process newly-encountered words than upon contextual originality per se. It is argued that both the involvement effect and the context effect may contribute to L2 vocabulary learning collaboratively and individually.

This study offers implications for L2 vocabulary learning through reading. Different sentence output tasks could be frequently deployed to facilitate vocabulary learning as long as learners have arrived at a level of L2 proficiency or vocabulary size sufficient to complete them. L2 teachers are encouraged to choose a sentence output task most appropriate for pedagogical goals. If the main teaching goal is to develop learners' translation skills with word learning as a by-product, for instance, L2-L1 translation exercises would well be a good choice. If developing learners' writing skills is the main teaching goal, sentence paraphrasing or writing could be designed to strengthen the link between new L2 word form

and meaning. Since one week after the initial encounters with the new words witnessed the learners' great loss of the word meaning knowledge, as found in this study, L2 teachers are informed that the learners should be offered multiple exposures to new words for successful learning [24, 31]. It is also suggested that multiple word encounters should be coupled with output tasks, as subsequent word retention was more contingent on elaborative processing of form–meaning relationships than on mere word frequency [10]. As Rott, Williams, and Cameron (2002) [34] conclude, "tasks that require the learner to elaborate the input and generate associations with prior experience and knowledge are considered ideal for promoting transfer of new knowledge to long-term memory" (p. 209).

This study made initial efforts to address the issue of contextual originality with regard to involvement loads and thus task effectiveness. The evidence in favor of the ILH is largely inconclusive with respect to initial word learning, further research should collect more data in the hope of finding compelling evidence for either the presence or absence of task effects. Research is still needed to examine the interplay between the task-induced involvement loads and the leaner-generated context in contributing to L2 vocabulary learning. Research may also be needed to quantify the degree of context generation just as Laufer and Hulstijn (2001) [23] did the degree of involvement in order to better understand the role of context in L2 vocabulary learning.

Notes

1. The analysis was conducted by implementing the program "anovaBF" from the R package "BayesFactor" developed by [26].
2. The statistics in the left panel of Figure 1 were used to compute Cohen's *d*s [5].
3. The statistics in the right panel of Figure 1 were used to compute Cohen's *d*s.
4. The Gelman-Rubin statistic (Rhat) is a convergence statistic when multiple chains are generated in parallel. Ideally, the Rhat statistic is 1.0 if the chains are fully converged; that is, the sampling distribution has fully converged to the posterior distribution. In general, an Rhat close to 1 (e.g., ± 0.1) suggests adequate convergence of the chains.

References

[1] Bao, G. (2015). Task type effects on English as a Foreign Language learners' acquisition of receptive and productive vocabulary knowledge. *System*, *53*, 84–95. https://doi.org/10.1016/j.system.2015.07.006

[2] Bao, G. (2019). Comparing input and output tasks in EFL learners' vocabulary acquisition. *TESOL International Journal*, *14 (*1), 1–12.

[3] Bao, G. (2019). The Bayesian approach to data analysis in applied linguistics. *Foreign Languages Research*, *176* (4), 16–23.

[4] Boers, F., & Lindstromberg, S. (2008). How cognitive linguistics can foster effective vocabulary teaching. In F. Boers & S. Lindstromberg (eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology* (pp. 1-64). Berlin: Mouton de Gruyter.

[5] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[6] 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. https://doi.org/10.1016/S0022-5371 (72)80001-X

[7] 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. https://doi.org/10.1037/0096-3445.104.3.268

[8] Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71* (9), 1–25. https://doi.org/10.18637/jss.v071.i09

[9] Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*, 207–218. https://doi.org/10.3758/s13423-017-1266-z

[10] 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. https://doi.org/10.1177/1362168811431377

[11] Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40* (2), 273–293. https://doi.org/10.2307/40264523

[12] Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or ILH: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research*, *47* (4), 859–869. https://doi.org/10.1007/s10936-018-9568-5

[13] Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL*, *41*, 87–106. https://doi.org/10.1515/iral.2003.007

[14] Hu, H.-c. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: ILH versus technique feature analysis. *System*, *56*, 28–39. https://doi.org/10.1016/j.system.2015.11.001

[15] Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the ILH in vocabulary acquisition. *Language Learning*, *51* (3), 539–558. https://doi.org/10.1111/0023-8333.00164

[16] 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. https://doi.org/10.1177/1362168808089922

[17] Kim, Y.-J. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58* (2), 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

[18] Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. San Diego, CA: Elsevier Inc.

[19] Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1* (2), 270–280. https://doi.org/10.1177/2515245918771304

[20] Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

[21] Laufer, B. (2017). From word parts to full texts: Searching for effective methods of vocabulary learning. *Language Teaching Research*, *21* (1), 5–11. https://doi.org/10.1177/1362168816683118

[22] 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. https://doi.org/10.1093/applin/amn018

[23] Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22* (1), 1–26. https://doi.org/10.1093/applin/22.1.1

[24] Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research*, *19* (6), 687–711. https://doi.org/10.1177/1362168814559797

[25] Lu, M. (2013). Effects of four vocabulary exercises on facilitating learning vocabulary meaning, form, and use. *TESOL Quarterly*, *47* (1), 167–176. https://doi.org/10.1002/tesq.79

[26] Morey, R. D., & Rouder, J. N. (2018). *Package "BayesFactor."* Accessed 10 March 2020 at https://cran.r-project.org/web/packages/BayesFactor/index.html

[27] Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle.

[28] Nist, S. L., & Mohr, C. (2002). *Advancing vocabulary skills*. Marlton, NJ: Townsend Press.

[29] Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68* (4): 1032–1075. https://doi.org/10.1111/lang.12310

[30] Norouzian, R., de Miranda, M., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *The Modern Language Journal*, *103* (1), 248–261. https://doi.org/10.1111/modl.12543

[31] Pichette, F., De Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics*, *33* (1), 66–82. https://doi.org/10.1093/applin/amr037

[32] Rassaei, E. (2017). Effects of three forms of reading-based output activity on L2 vocabulary learning. *Language Teaching Research*, *21* (1), 76–95. https://doi.org/10.1177/1362168815606160

[33] R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Accessed 5 January 2019 at https://www.R-project.org/

[34] Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input–output cycles on lexical acquisition and retention. *Language Teaching Researc*h, *6* (3), 183–222. https://doi.org/10.1191/1362168802lr108oa

[35] Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing*, *18* (1), 55–88. https://doi.org/10.1177/026553220101800103

[36] Yanagisawa, A., & Webb S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, 2021, *71* (2), 487–536. https://doi.org/10.1111/lang.12444

[37] Yeung, S. S., Ng, M. L., & King, R. B (2016). English vocabulary instruction through storybook reading for Chinese EFL kindergarteners: Comparing rich, embedded, and incidental approaches. *The Asian EFL Journal*, *18* (2), 81–104.

[38] Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, *21* (1), 54–75. https://doi.org/10.1177/1362168816652418