OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 14, Number 1 (2025)
https://doi.org/10.29140/vli.v14n1.102482

Castledown

Generation and L2 Vocabulary Learning: A Classroom Action Study on the Efficacy of the Generation Desirable Difficulty in Learning L2 Vocabulary

John Duplice
Associate Professor
Department of Business Administration, Meiji University
duplice@meiji.ac.jp

Abstract

The present study investigates the use of generative strategies and their effect on second language (L2) vocabulary retention and learner confidence in vocabulary knowledge over time. Generative strategies (or generation) are one in a group of learning methods known as desirable difficulties – strategies that improve long-term learning by adding challenge to the learning process (Bjork & Bjork, 2011). Previous research has established the effectiveness of other desirable difficulties, such as spacing, retrieval, and interleaving, in L2 vocabulary acquisition (Kang et al., 2013; Lafleur & Kanazawa, 2024; Nakata, 2015). However, the effect generative strategies have on L2 vocabulary learning over shorter timeframes (e.g., a few days to a week) has been found to lack efficacy (Barcroft, 2009). Nevertheless, long-term vocabulary outcomes implementing this strategy in L2 learning remain unclear. To address this gap, this study examined the effectiveness of generation tasks used as the sole desirable difficulty for vocabulary recall over two intervals: three weeks and 15 weeks. Using quantitative methods in a classroom action research project, this study investigated 38 first-year Japanese university English language learners at the

Data Availability Statement: All relevant data is available at https://osf.io/2vcg9/

B1 to B1+ CEFR English proficiency level. The treatments included generation tasks in the form of generating novel sentences for half of the target vocabulary and a control treatment utilizing other non-generative methods for the other half through two separate lessons. Additionally, pre and post-tests were implemented to assess the perceived confidence and actual knowledge of the target vocabulary. The study`s findings on longer-term timeframes were not markedly better than those of shorter timeframes and only showed slight improvement. While the study failed to demonstrate a clear benefit for the additional effort required by generation tasks, it did not identify significant costs from using generation on the specific population studied, as found in other studies. Implications for the findings and future research are included.

Introduction and Background

The integration of second language (L2) vocabulary instruction and the capacity to recall learned terms beyond immediate assessments are crucial for L2 learners. One notably effective pedagogical strategy involves using desirable difficulties—techniques introducing challenges to promote long-term retention (Bjork & Bjork, 2011). Among the strategies recognized as desirable difficulties in L2 vocabulary acquisition are spaced repetition, interleaving, and retrieval practices. Another particularly challenging type of desirable difficulty is the generative or generation strategy, which requires deeper engagement with the material, resulting in a deliberate slowing down of studying during a study session. This approach compels learners to construct meaning and relate new information to their pre-existing knowledge or schema (Enser & Enser, 2020; Weinstein et al., 2018). Generation is often implemented through elaboration activities or summarization tasks across various contexts, including reading and writing in L1, life and physical sciences, and mathematics (e.g., Wittrock, 1974a). Research within the learning sciences demonstrates that this form of active involvement may enhance recall across different first language (L1) scenarios (Bertsch et al., 2007; McCurdy et al., 2020), while less interactive non-generative methods, such as highlighting and rereading, provide minimal to no advantage (Dunlosky et al., 2013). In the present study`s L2 vocabulary learning context, generating novel sentences through connecting the target word by generating a specific connection with the individual student requires the learner to write an original sentence and take time to think of a specific situation the learner relates to. The student must then include multiple aspects (e.g., new L2 vocabulary, grammar structure of the word for the sentence, meaning related to the student, and expressing this all in a coherent sentence or two). The specifics of how generation was implemented in this study are reported in more detail in the procedure section.

The effectiveness of generation is linked to its elevated cognitive demand and its association with the learner's schema, which is thought to facilitate the learning process (Bjork & Bjork, 2011; Roediger & Karpicke, 2006). While generation benefits have been researched more thoroughly in L1 contexts with some positive outcomes, generation`s effectiveness, or lack thereof, in L2 contexts is less clear. For instance, limited research investigating the generation application in L2 vocabulary learning has

not yielded the same advantages as in L1 contexts. Barcroft (2009) found that when generation is employed as the sole study method, such as creating novel meanings through synonyms, it results in lower L2 vocabulary recall in the short term (three days and one week). In a subsequent study, Barcroft (2015) examined generation in L2 vocabulary acquisition by offering scaffolding that included target vocabulary and definitions before the generation task and a retrieval activity (another desirable difficulty). By reducing the difficulty of generation, thereby tempering the argument for its effectiveness, and including retrieval tasks, Barcroft posits that generation may only be advantageous for L2 vocabulary acquisition if its difficulty is lowered and not utilized as the sole desirable difficulty. Moreover, the improvement in Barcroft`s 2015 study could be solely due to using retrieval despite the incorporation of generation. This is also discussed outside of L2 specific contexts. Dunlosky (2013), looking at learning in education in general (L1 and L2 contexts), argues that the lack of clear benefit and the considerable effort required by the learner through using generation call into question its efficacy as a sole learning method. Additionally, given that generation encompasses various factors (e.g., personalization of generated material and grammar use) and is related to other desirable difficulties (e.g., retrieval), the benefits in all timeframes are less clear than those of other desirable difficulties. The discourse surrounding efficacy discussed here addresses the feasibility of generation for L2 learners. To date, the research focus has predominantly been on immediate and short-term outcomes, finding a lack of efficacy in these timeframes. What remains less understood is whether there is a long-term effectiveness of generation in L2 vocabulary learning (e.g., over three weeks or more) not seen in the short and intermediate timeframes, which is essential for acquiring L2 vocabulary for sustained use and knowledge transfer.

Study Introduction and Rationale

This study investigated the impact of generation on long-term retention of L2 vocabulary and learners' confidence in their knowledge of the target vocabulary. While the benefits of spacing, retrieval, and interleaving are well established in L2 vocabulary acquisition (e.g., Barcroft, 2007; Karpicke & Roediger, 2008; Nakata, 2011; 2017), the efficacy of generation remains less understood. Therefore, this study focused on generation over two periods (three and 15 weeks) post-initial exposure through classroom action research conducted with first-year English language learners (ELLs) at a Japanese university. Confidence was included to assess perceived knowledge versus actual knowledge of vocabulary. As there are few studies on confidence beyond Pellicer-Sanchez et al. (2021), this aspect of the study is more exploratory in nature than the knowledge aspect.

Previous research has shown positive effects of generation tasks on long-term learning (e.g., Jacoby & Chestnut, 1978) through L1 summarization (e.g., Wittrock, 1974b; 1989), reading tasks (e.g., DeWinstanley & Bjork, 2004), and L1 word lists (e.g., McDaniel et al., 1988). However, some researchers argue generation may cause an increase in cognitive load from the inherit complexity of generating content due to the need to use multiple aspects or skills (Chen et al., 2015; 2016; 2018). In the L2 context, the complexity is likely increased from using L2 (e.g., retrieval, word form use, L2 grammar). Cognitive load is the amount of information a learner can manage

in their limited working memory (Sweller, 1988). When the demands on attention surpass these limits, it can result in cognitive overload, hindering effective new learning, a counterargument to desirable difficulty use during the encoding process (Cadaret & Yates, 2021; Chen et al., 2015; 2016; 2018).

Methodology

The methodology of this study was grounded in a predetermined curriculum within a classroom action research framework, involving university-assigned students as participants. Due to the constraints of convenience, the options for methodology were limited, impacting the number of participants and the timeframe for post-tests to assess the confidence and knowledge of the target vocabulary, which is what was specifically investigated in this study.

What it means to know a word can vary depending on what is being assessed, and, in many L2 classroom situations, "word knowledge" is assessed by simply recalling the provided definitions in the textbook. However, word knowledge encompasses many aspects, such as word meanings, collocations, semantic relationships, and syntactic patterning (Cobb, 1999; Webb, 2020). The present study required participants to produce their own definitions and novel context-based examples by writing personalized sentences using the target word, minimizing reliance on guessing through word forms or multiple-choice formats. Additionally, the research examined learners' confidence in their vocabulary knowledge, which may not align with their actual understanding. This discrepancy could lead to false fluency (a false sense of ability), especially when knowledge stems from similarly spelled words or contexts. Building on findings by Pellicer-Sanchez et al. (2021) that demonstrated a relationship between accuracy and confidence in both L1 and L2 learners regarding word collocations, this study evaluated participants' confidence in target words and compared these ratings to their actual usage ability.

The vocabulary items were selected from the lessons in Nation (2018) *4000 Essential English Words 5 Second Edition*, a publicly available L2 vocabulary learning textbook. A pre-test of the same vocabulary used in the post-tests was administered at the study's onset for two purposes. The first was to establish a self-reported participant confidence score, termed "Confidence," reflecting the learners` assurance in using each vocabulary item. Participants rated their knowledge of each word on a four-point scale (1: do not know; 2: seen before but do not know the meaning; 3: understands in context but cannot use; 4: understands and can use confidently), similar to scales used by Pellicer-Sanchez et al. (2021).

Secondly, the pre-test assessed whether students understood and could use the vocabulary items. Following their self-reported scores, participants defined each word and provided an example of comprehension, establishing the "Knowledge" score. Responses were marked as either correct or incorrect, according to the participant`s ability to clearly provide the vocabulary definition and a proper example. Figure 1 shows an example of a model correct answer. Since the lesson had not been taught yet, correct usage in any context was acceptable; however, context-specific accuracy was required in future post-tests that followed the lesson and the definition of how the

target vocabulary terms were defined in the lesson`s context. Partially correct responses were excluded from the analysis, and a simplified correct/incorrect scoring method was employed to indicate knowledge improvement due to the treatment, avoiding the complexity of vague or guessed answers. The study did not include follow-up activities beyond the initial vocabulary introduction, and verifying participants did not partake in independent studying was not possible. However, instructions were given at each meeting, emphasizing the importance of not studying the materials outside the classroom. Furthermore, participants were told the post-test results would not affect their grades. To minimize external influences, tasks were completed in class, with materials collected by the instructor.

Data from pre- and post-tests were collected over 15 weeks during face-to-face classes, allowing for compliance monitoring. The three-week post-test timing aligned with Nation's (2008) and Schmitt's (2010) recommendations for assessing durable vocabulary acquisition. The second post-test, administered 15 weeks after initial exposure, provided a longer interval than reported in previous L2 vocabulary studies on generation use by Barcroft (2009) and Sun and Fang (2022). Long-term recall was evaluated at three weeks and again at 15 weeks, reflecting realistic L2 vocabulary retention for long-term use and learning transfer. The 15-week post-tests also considered the potential decline in language engagement over an eight-week summer break, necessitating the extended assessment period. All data from the two groups were compiled into one file.

Participants

The study initially comprised 38 first-year university students in Japan (L1 Japanese), averaging 18 years of age, typical for this demographic. Participants were L2 speakers of English at the B1 to B1+ Common European Framework (CEFR) level. Prior to the academic year, they completed the Test of English for Academic Purposes (TEAP), which the university used for class placement based on estimated English proficiency. The TEAP evaluated reading and listening skills, however, it did not assess writing or speaking. Participants' majors varied (e.g., global studies, journalism), with none specializing in English or English literature.

This vocabulary study included two groups (Group A = 16, Group B = 22) with 38 initial participants. Data from 15 participants (Group A = 8, Group B = 7) were excluded due to incomplete data caused by multiple class absences primarily related to COVID-19. The final analysis included data from 23 participants (6 males, 17 females: Group A males = 2, females = 6; Group B males = 4, females = 11) who completed all study components. Both groups attended compulsory Academic Communication classes bi-weekly over two semesters. Classes for Group A occurred on Mondays and Thursdays from 9:00 to 10:40, while Group B met on Mondays and Thursdays from 15:25 to 17:05. Participants received no compensation, and test scores did not contribute to course grades; however, attendance was mandatory, influencing participation grades. A consent form was provided, allowing participants to opt out at any time by contacting the instructor, and this study was approved by the University Research and Ethics Committee.

Study Materials

The study materials for each lesson included a pre-test (Appendices 2a & 2b) and paper-based lessons of 7–8 pages, comprising 20 target vocabulary items with definitions, images, sample sentences, a 270-word reading activity incorporating the vocabulary, and a post-reading comprehension task.

Rationale for Target Vocabulary Chosen

The vocabulary level of the target terms was chosen so that they would be challenging yet accessible for B1 to B1+ CEFR level. Furthermore, academic vocabulary was chosen so that students would be less likely to encounter them daily. These were provided in Nation (2018) *4000 Essential English Words 5 Second Edition*, which included pre-selected B2-level vocabulary, enhancing exposure to academic terms. The vocabulary items also came from Coxhead's (2000) Academic Word List (AWL), representing words estimated to be above the participants' level and not commonly used outside school. Each lesson contained 20 vocabulary words (40 total), ensuring balanced difficulty across treatments. The 10 words chosen for the control treatment were included in tasks provided in the ELL textbook for convenience, while the other tasks that included the generation treatment words were not used and replaced with the generation treatment. After the pre-test, different treatments were implemented to modify how participants learned vocabulary. Lessons were distributed right after the pre-test, with all references to the textbook and author removed to prevent independent identification by participants.

Procedure

This study spanned 15 weeks, with both groups completing identical lessons and vocabulary items. The groups were separated to consider different class dynamics. Both groups received the same baseline and generation vocabulary, with participants encountering half of the words in each condition. Two treatments were administered: a baseline control condition involving non-generative tasks (e.g., vocabulary-to-definition matching) and a generation task where participants crafted novel sentences using target vocabulary. The use of bilingual dictionaries was allowed, but example sentences from the dictionaries were not. Table 1 outlines the corresponding groups and treatments for each lesson, along with their pre- and post-tests.

The procedure included:

1. A pre-test of 20 vocabulary items from "The History of Chocolate" (collecting Confidence and Knowledge scores).
2. An introductory lesson covering vocabulary definitions and sample sentences, followed by comprehension questions.
3. A second part with generation tasks for ten items and baseline tasks for the other ten.

In Week Two, participants received specific instructions similar to Week One, and both groups completed the lesson on the same day but in different class periods. Each

Table 1 *Week and Topic for each Group, Treatment and Post-Tests*

Week (topic)	Groups A & B (Vocabulary conditions regarding treatment and specific words)
Week 1 (The History of Chocolate) Pre-test + Treatment	10 vocabulary items –Generation treatment 10 vocabulary items –Baseline treatment 20 vocabulary items in total
Week 2 (How Did the Dinosaurs Really Die?) Pre-test + Treatment	10 vocabulary items –Generation treatment 10 vocabulary items –Baseline treatment 20 vocabulary items in total
Week 4 (The History of Chocolate)	Post-test 1 (20 vocabulary items per post-test)
Week 5 (How Did the Dinosaurs Really Die?)	Post-test 1 (20 vocabulary items per post-test)
Week 15 (The History of Chocolate) Post-test 2	Post-test 2 (20 vocabulary items per post-test)
Week 15 (How Did the Dinosaurs Really Die?) Post-test 2	Post-test 2 (20 vocabulary items per post-test)

session lasted 100 minutes, providing ample time for the activities. Baseline activities consisted of fill-in tasks, synonym comparisons, definition matching, and a crossword puzzle to balance time for generation tasks, with both conditions taking around 40 minutes. Participants created two to three personalized sentences for each vocabulary item in the generation task based on a provided example (see Figure 1).

Participants completed lessons individually in the physical classroom in front of the instructor using pencil and printouts provided, after which the instructor verified each student`s completion of baseline activities and generation sentences following the example provided and using novel sentences specifically related to the individual student. The instructor ensured that the sentences had sufficient content and accurate

For example:
procrastination*: It means not doing something immediately even though it should be done right away.*

Procrastination is a problem for me at school. Last week I waited until just before the deadline for my English essay before starting it. I was up all night writing it, and I did not get a good grade because I **procrastinated.**

Figure 1 *Model Answer at the Top of the Generation Lesson Task.*

usage before students submitted their work. In the baseline activities, the focus was on confirming that students used the target vocabulary correctly rather than guessing answers. Once verification was completed by the instructor using an answer sheet for the baseline answers, participants formed groups of three to compare their baseline answers and share their generation sentences. Groups received correct answers for the baseline tasks and reading comprehension questions after sharing. At the end of each session, the instructor collected lesson materials, with no further exposure to vocabulary items provided until three weeks later during post-test 1, which mirrored the pre-test in assessing Confidence and Knowledge scores. The second lesson, "How the Dinosaurs Really Died," followed the same structure one week after the first lesson, with post-test 1 occurring three weeks post-lesson.

The second post-tests, one for each topic, were administered 15 weeks after vocabulary introduction, occurring one week apart. This timing ensured that the intervals of lessons and post-test 2 (15 weeks from the lesson and 12 weeks from post-test 1) were equal between groups. Post-test 2 matched the pre-test and post-test 1 format, with students providing Confidence and Knowledge ratings for each word assessed.

Participants completed the pre and post-tests within thirty minutes in the instructor's presence, which ensured there was enough time to finish. Participants were instructed not to refer to materials and to rely solely on memory. Given that self-reported Confidence scores on vocabulary knowledge were included, participants needed to understand that lower scores would not impact their grades. Therefore, students were again informed that the scores would not affect their class grades.

Data Collection and Assessment

Following vocabulary test procedures laid out by Nation and Webb (2011), pre and post-tests maintained identical questions to eliminate discrepancies in test difficulty. The tests were conducted entirely in English, and there was no Japanese in the instructions. Adjacent to each test item, participants recorded their self-reported knowledge of vocabulary, with a reference scale provided at the top of the test (e.g., 3, indicating an understanding of the word in context, but not its usage).

For Knowledge assessments, participants had space for detailed responses, allowing multiple sentences to explain meanings or provide contextual examples. The instructor evaluated answers as correct or incorrect, permitting spelling and grammar errors as long as they did not obscure comprehension of meaning or usage. Answers deemed vague or unclear were rated as incorrect, as were interpretations of vocabulary differing from the lesson content. An example of a correct answer for the word *genetic* was *"Genetic is something from your parents or family that you can the same. Almost people believe that heredity of character in family. My family too, my dad and my face is very similar. I think that is genetic effect."* While the grammar is not perfect, the ability to show a relationship between genetics and family, along with a clear novel set of sentences personalizing the word, would make this a correct answer. However, if the sentence was not personalized or too vague to clearly show understanding (e.g., *I have the genetics. Lots of people have genetics and they are very common in Japan.)*, it was marked as incorrect.

Results

To compare the effects of two treatment conditions, two linear models were constructed using Jamovi Project version 2.3, R version 4.1, and GAMLj (Gallucci, 2019). A linear mixed model analyzed Confidence scores (1–4) for vocabulary items, treating participants and specific items as random variables to account for intra-participant effects. For Knowledge scores, a generalized linear model with binomial distribution was used, marking answers as correct or incorrect. Statistical analyses were conducted using NHST to evaluate the effectiveness of the treatments. Results include 95% Confidence Intervals to provide a range of plausible values for the observed effects, thereby in order to convey a degree of uncertainty inherent in the estimates.

Each vocabulary item's Confidence and Knowledge scores were treated as individual data points across pre and post-tests. Confidence scores ranged from "1" (no knowledge) to "4" (full confidence), while Knowledge scores were marked as "1" for correct and "0" for incorrect. This resulted in 60 data points per participant for the pre-test and two post-tests in Confidence scores, and 60 data points in Knowledge scores (120 total for each lesson). The combined pre- and post-tests yielded 240 data points. With 23 participants, there were 2,760 data points (23 participants x 120 data points) for each test type (Confidence and Knowledge).

Results in Aggregate for Confidence Scores

The combined results of the participants' self-reported confidence levels are displayed in Table 2. The scores vary from a minimum of "1" (indicating no familiarity with the vocabulary item) to a maximum of "4" (indicating certainty in understanding and using the vocabulary).

A preliminary model incorporated fixed effects for condition, with treatment coded using the baseline as the reference level, and for test type, which was coded with the pre-test as the reference level. Additionally, the model featured random intercepts for both subjects and vocabulary items, as well as by-subject slopes for condition and by-item random slopes for the impact of test type (covering pre- and post-tests). This approach aimed to enhance the random effect structure and thus improve the model's generalizability (Barr et al., 2013). The study analyzed self-reported Confidence scores across all three tests (pre-test, post-test 1, post-test 2). The treatment's overall effect was not found to be statistically significant ($\beta = 0.06$, $t = 0.50$, $p = .624$). The 95% CI

Table 2 *Mean Confidence Scores by Treatment (and test type with standard deviation (SD) and 95% Confidence Intervals (CIs)*

	Baseline condition			Generation condition		
	Pre-test	Post-test 1	Post-test 2	Pre-test	Post-test 1	Post-test 2
Mean	3.08	3.43	3.41	3.09	3.55	3.47
SD	1.002	0.859	0.755	0.961	0.772	0.733
95% CIs	2.99, 3.17	3.36, 3.51	3.34, 3.48	3.00, 3.17	3.48, 3.62	3.41, 3.54

scores ranged from –0.187 to 0.313. There were significant simple effects observed for post-test 2 (t = 6.27, p < .001, 95% CI 0.245 to 0.468), while post-test 1 did not show significance (t = 1.68, p = .106, 95% CI –0.042 to 0.549), indicating that the later post-test reflected a marked improvement over pre-test scores, unlike the first post-test. The interactions of treatment with post-test 1 (t = 1.15, p = .259, 95% CI –0.079 to 0.301) and post-test 2 (t = 0.52, p = .605, 95% CI –0.144 to 0.248) also did not yield significant results (refer to Table 3). Post Hoc pairwise analyses were conducted using the GAMLj package, implementing Bonferroni correction for multiple comparisons, which revealed no significant differences in scores between the generation condition and the pre-test (z = 0.34, p = 1.00), at post-test 1 (z = –2.62, p = .132), or at post-test 2 (z = –1.64, p = 1.00). Figure 2 illustrates the treatment comparisons at each testing point.

Table 3. *Treatment interactions t and p-values with confidence intervals included*

				95% Confidence Interval				
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	3.2902	0.0912	3.1116	3.469	51.8	36.094	< .001
Treatment1	g – b	0.0630	0.1277	-0.1872	0.313	38.8	0.494	0.624
Test_type1	Post-test1 – Pre-test	0.2533	0.1508	-0.0423	0.549	24.4	1.680	0.106
Test_type2	Post-test2 – Pre-test	0.3565	0.0569	0.2451	0.468	34.5	6.270	< .001
Treatment1 * Test_type1	g – b * Post-test1 – Pre-test	0.1109	0.0969	-0.0790	0.301	42.0	1.145	0.259
Treatment1 * Test_type2	g – b * Post-test2 – Pre-test	0.0522	0.1001	-0.1440	0.248	40.5	0.521	0.605

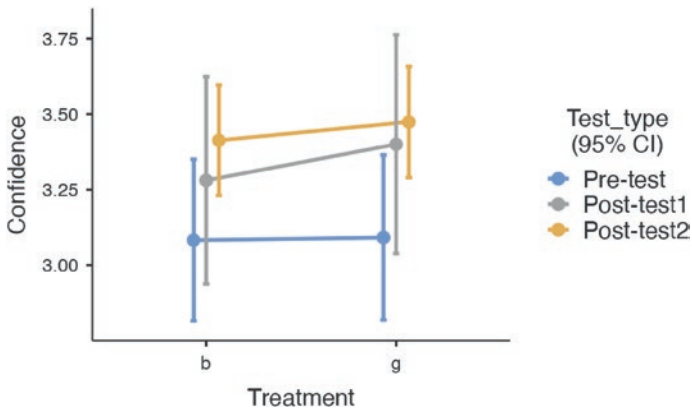

Figure 2 *Side-by-Side Comparison of Confidence Ratings by Treatment (B = Baseline, G = Generation) Including 95% Confidence Intervals (CI).*

Results in Aggregate for Knowledge Scores

The overall results of Knowledge scores are shown in Table 4, illustrating the percentage of vocabulary items correctly answered.

A generalized mixed model was fitted with treatment-coded fixed effects of treatment (reference = Baseline condition) and test type (reference = pre-test). Random intercepts were included along with by-subject slopes for condition and test type, and by-item slopes for test type since one item was never seen in different conditions. This was included to maximize the random effect structure and, therefore, the generalizability of the model (Barr, Levy, Scheepers, & Tily, 2013). The overall effects of treatment ($\beta = -0.19$, $z = -0.44$, $p = .661$) were found not to be significant. The 95% CI scores ranged from 0.350 to 1.95There were simple effects of post-test 1 ($z = 3.95$, $p = <.001$), but not post-test 2 ($z = 1.61$, $p = .108$, 95% CI 0.933 to 2.03), showing that the first post-test (but not the second) showed a significant improvement over

Table 4 *Mean Knowledge Scores for Items Answered Correctly by Treatment, Test Type with Standard Deviation (SD) and 95% Confidence Intervals (CIs)*

	Baseline condition			Generation condition		
Tests	**Pre**	**Post-1**	**Post-2**	**Pre**	**Post-1**	**Post-2**
Mean	0.40	0.52	0.46	0.36	0.59	0.51
SD	0.490	0.500	0.499	0.479	0.492	0.500
95%, Cis	0.353, 0.443	0.570, 0.500	0.415, 0.507	0.313, 0.400	0.548, 0.639	0.467, 0.559

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

Table 5 *Treatment Interactions z and p-Values with Confidence Intervals Included*

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	-0.600	0.329	0.549	0.288	1.05	–1.824	0.068
Treatment1	g – b	-0.192	0.438	0.825	0.350	1.95	–0.438	0.661
Test_type1	Post-test1 – Pre-test	0.682	0.173	1.979	1.411	2.78	3.954	< .001
Test_type2	Post-test2 – Pre-test	0.319	0.198	1.376	0.933	2.03	1.609	0.108
Treatment1 * Test_type1	g – b * Post-test1 – Pre-test	0.734	0.280	2.083	1.204	3.60	2.624	0.009
Treatment1 * Test_type2	g – b * Post-test2 – Pre-test	0.620	0.247	1.860	1.147	3.02	2.515	0.012

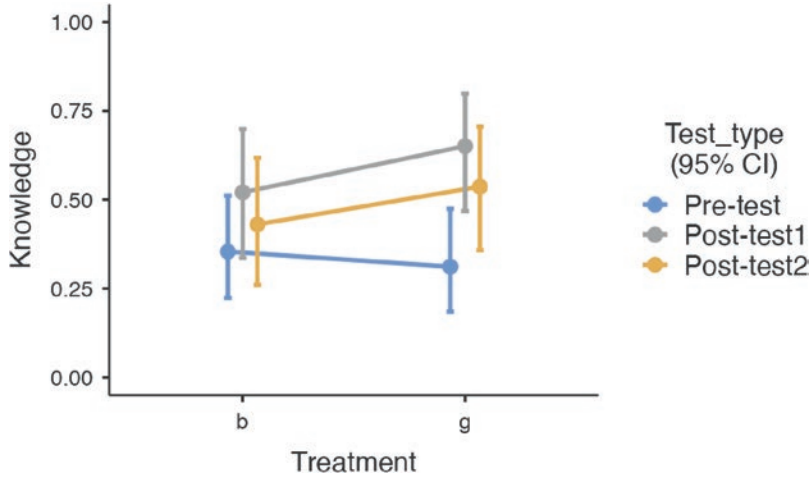

Figure 3 *Interaction of Treatment (b = Baseline, g = Generation) and test type for Knowledge Scores with 95% Confidence Intervals.*

the pre-test scores. Interactions of treatment with post-test 1 (z = 2.62, p = .009, 95% CI 1.204 to 3.60) and post-test 2 (z = 2.56, p = .012, 95% CI 1.147 to 3.02) showed a significant effect (see Table 5). Interactions show an overall increase in scores from the pre-test for generation treatment over the control. However, Post Hoc pairwise analyses with Bonferroni correction for multiple comparisons showed there were no significant effects of treatment at each test point: pre-test (z = 0.44, p = 1.00), post-test 1 (z = –1.85, p = .972), or post-test 2 (z = –1.03, p = 1.00) as shown in Figure 3.

Discussion

In this research study, data was gathered on participants' perceived knowledge (Confidence scores) and actual knowledge (Knowledge scores) relating to newly acquired L2 vocabulary. The collected data included scores from a pre-test, post-test 1, and post-test 2 for both measures. The results indicated that neither perceived nor actual knowledge demonstrated a significant advantage from the generation activity compared to baseline activities; however, there was no adverse effect from generation either, indicating no detriment compared to the baseline. While the straightforward comparisons did not yield statistically significant results, as shown by the Post Hoc tests at each testing phase, the interactions in the Knowledge scores, illustrated in Table 5 and Figure 3, suggest a potential advantage for the generation condition from the pre-test to the post-tests. The generation treatment did not negatively influence recall for both Confidence and Knowledge scores. Barcroft (2009) noted that generation was ineffective for L2 vocabulary recall when it served as the primary design demand. Similarly, this study does not present a clear overall advantage. However, the significant interactions regarding condition and test timing in Knowledge scores imply that the increase in the generation condition may have outpaced that in the control condition. In contrast to Barcroft (2009), there was no detrimental impact noted from generation

in this study. Additionally, there were notable differences between Barcroft's research and the current study. One key distinction is that Barcroft focused on synonym generation, while the present study involved the generation of novel sentences. Synonyms may lack the same personal relevance for learners compared to the creation of novel sentences. Another critical difference pertains to the participants; Barcroft's subjects were Spanish L1 speakers learning English as a second language, likely having more cognates than the Japanese L1 speakers in the present study. Lastly, the timing of vocabulary exposure varied significantly, as Barcroft considered immediate post-test results after the lesson, which may reflect performance, an observable outcome during or immediately after studying, rather than actual learning, which is considered relatively permanent (Soderstrom & Bjork, 2015). The present study aimed at long-term recall measured at three weeks and 15 weeks after initial exposure to the target vocabulary items to study the effect the generative treatment has on long-term learning.

In a smaller study more closely related to the current research, Iwata (2012) focused on Japanese L1 high school students learning English L2 vocabulary through sentence generation. Iwata's findings indicated a significant improvement in the generation group over the control group at both one-week and four-week post-test intervals. The control group in Iwata's study engaged in sentence translation from English to Japanese, embedding the target vocabulary contextually. The results of the present study fall in between those of Barcroft (2009) and Iwata (2012), highlighting the necessity for further research with a larger sample size, as the findings suggest a significant interaction, indicating there may be a potential benefit of generation.

Limitations

This study has notable limitations affecting the results. Firstly, the small sample size may not accurately reflect the broader population and did not allow for gender to be studied as a covariate. A larger sample size and more diverse population may provide more conclusive findings. Although pairwise differences at each test point were insignificant, Figures 2 and 3 suggest a trajectory of increased recall in the generation treatment, which might reach significance with a larger sample, but this cannot be determined with the current data. The confidence intervals suggest that while generation did not improve vocabulary recall in this study, the exact magnitude of the generation effect may vary, as illustrated in the Confidence Interval range. This variability underscores the importance of considering contextual factors when applying these results to different settings. Secondly, the study focused exclusively on first-year Japanese university students in Tokyo, limiting the generalizability of the findings to other L2 vocabulary learners. Additionally, the research utilized only one measure of vocabulary knowledge—recall—accounting for word explanation and sentence generation, which is a limitation since it did not consider other factors like interlexical and intralexical aspects. Finally, student answers were only conducted by a single rater.

Conclusion

Limited time and resources in L2 classrooms necessitate effective teaching methods to help students achieve their goals efficiently. This study examined the role of generation

in learning L2 vocabulary, which is crucial for second language acquisition (Schmitt & Schmitt, 2020). While the study did not find clear benefits from generation, it also did not replicate the costs noted by Barcroft (2009; 2015). Consequently, given the additional effort and time required for generation tasks, this strategy may not be optimal for L2 vocabulary acquisition. Alternative approaches, such as spacing, interleaving, and retrieval when used as the primary strategy, may yield better results when encoding new L2 vocabulary.

Findings suggest that generation advantages observed in some L1 contexts may not translate effectively to L2 settings during initial encoding. Recent research indicates that generation may serve best as a consolidation strategy post-encoding (Roelle et al., 2022a; 2022b; Schindler & Richter, 2023). Despite its potential benefits for reinforcing vocabulary retention, further investigation is warranted. One possible approach includes providing clear contextual examples of target terms during initial exposure, followed by retrieval practice (e.g., flashcards), and incorporating spacing and interleaving until the vocabulary is embedded in long-term memory. Once the learners have achieved this, they may effectively generate and transfer knowledge of the L2 vocabulary to new contexts. By initially implementing these less complex desirable difficulties, cognitive load may be minimized, enhancing the learning experience. The present study's choice to prioritize generation—an inherently more complex strategy—over simpler methods did not yield the expected enhancement in recall over the longer period of time investigated, indicating that less demanding strategies may be more effective during initial vocabulary acquisition.

References

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly*, *43*(1), 79–103. https://doi.org/10.1002/j.1545-7249.2009.tb00228.x

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*(59–68). https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/04/EBjork_RBjork_2011.pdf

Cadaret, C. N., & Yates, D. T. (2021). Homework questions designed to require higher-order cognitive skills in an undergraduate animal physiology course did not produce

desirable difficulties, testing effects, or improvements in information retention. *Journal of Animal Science*, *99*(9), skab246. https://doi.org/10.1093/jas/skab246

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable difficulty effects in the learning of high-element interactivity materials. *Frontiers in Psychology*, 1483. https://doi.org/10.3389/fpsyg.2018.01483

Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology*, *107*(3), 689. https://doi.org/10.1037/edu0000018

Chen, O., Kalyuga, S., & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction*, *45*, 20–30. https://doi.org/10.1016/j.learninstruc.2016.06.007

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, *12*(4), 345–360. https://doi.org/10.1076/call.12.4.345.5699

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Dunlosky, J. (2013). Strengthening the student toolbox: Study strategies to boost learning. *American Educator*, *37*(3), 12–21.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Enser, Z., & Enser, M. (2020). *Fiorella & Mayer's generative learning in action*. John Catt Educational.

Gallucci, M. (2019). *GAMLj: General analyses for linear models. [jamovi module]*.

Iwata, A. (2012). The effectiveness of summary-writing activities on the improvement of Japanese high school EFL students' writing abilities. *Research Bulletin of English Teaching*, 1–22.

Jacoby, J., Chestnut, R. W., & Fisher, W. A. (1978). A behavioral process approach to information acquisition in nondurable purchasing. *Journal of Marketing Research*, *15*(4), 532–544. https://doi.org/10.1177/002224377801500403

Kang, S. H., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, *20*, 1259–1265. https://doi.org/10.3758/s13423-013-0450-z

Karpicke, J. D., & Roediger III, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Lafleur, L., & Kanazawa, Y. (2024). The effects of interleaved spaced repetition learning on vocabulary knowledge. *Computer-Assisted Language Learning Electronic Journal*, *25*(3), 172–194.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*, 1139–1165. https://doi.org/10.3758/s13423-020-01762-3

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language*, *27*(5), 521–536. https://doi.org/10.1016/0749-596x(88)90023-x

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38. https://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, *39*(4), 653–679. https://doi.org/doi.org/10.1017/S0272263116000280

Nation, P. (2008). Lexical awareness in second language learning. *Encyclopedia of Language and Education*, *6*, 167–179. https://doi.org/10.1007/978-0-387-30424-3_147

Nation, P. (2018). *4000 Essential English Words 5 Second Ed.* Compass Publishing.

Nation, I. S., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning Boston, MA.

Pellicer-Sánchez, A., Vilkaitė-Lozdienė, L., & Siyanova-Chanturia, A. (2021). Examining L2 learners' confidence of collocational knowledge. *Vocabulary Theory, Patterning and Teaching*, *152*, 121. https://doi.org/10.2307/jj.22730558.12

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). *Linking cognitive science to education: Generation and interleaving effects*. *27*(27).

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022a). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction*, *80*, 101634. https://doi.org/doi.org/10.1016/j.learninstruc.2022.101634

Roelle, J., Schweppe, J., Endres, T., Lachner, A., von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., & Scheiter, K. (2022b). Combining retrieval practice and generative learning in educational contexts. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*. https://doi.org/10.1026/0049-8637/a000261

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press.

Schindler, J., & Richter, T. (2023). Text generation benefits learning: A meta-analytic review. *Educational Psychology Review*, *35*(2), 44. https://doi.org/10.1007/s10648-023-09758-w

Sun, H., & Fang, S. (2022). Paired-associate second language vocabulary learning: The role of L1 translation familiarity. *Journal of Asia TEFL*, *19*(1), 50. https://doi.org/10.18823/asiatefl.2022.19.1.4.50

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier. https://doi.org/10.1016/b978-0-12-387691-1.00002-8

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Webb, S. (2020). Approaches to learning, testing and researching L2 vocabulary. *Approaches to Learning, Testing and Researching L2 Vocabulary*, 1–240. https://doi.org/10.1075/bct.109

Weinstein, Y., Sumeracki, M., & Caviglioli, O. (2018). *Understanding how we learn: A visual guide*. Routledge.

Wittrock, M. C. (1974a). A generative model of mathematics learning. *Journal for Research in Mathematics Education*, 181–196. https://doi.org/10.2307/748845

Wittrock, M. C. (1974b). Learning as a generative process. *Educational Psychologist*, *11*(2), 87–95. https://doi.org/doi.org/10.1080/00461527409529129

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*(4), 345–376. https://doi.org/doi.org/10.1207/s15326985ep24042

Appendices

Appendix 1a: Pre / Post Assessments for Lesson 1

Target Vocabulary Assessment

The words below are the target vocabulary for this lesson. Give each word a score of 1, 2, 3, 4 for your current understanding of each word. If you know the word, please write a sentence using it or the meaning of the word in the blank space.

1) I do not know this word.
2) I have seen the word before, but I am not sure what it means.
3) I understand this word when I see or hear it in a sentence, but I don`t know how to use it.
4) I know this word and can use it in my speaking and writing.

Vocabulary Word	Score 1,2,3,4	Definition or sample sentence using the vocabulary word to show you can use it
aroma		
beverage		
cluster		
combine		
condensed		
contemporary		
cultivate		
divine		
humid		
odor		
palate		
paradise		
plantation		
rapid		
rate		
soothing		
subtle		
texture		
toxic		
vary		

Appendix 1b: Pre / Post Assessment for Lesson 2

Target Vocabulary Assessment

The words below are the target vocabulary for this lesson. Give each word a score of 1, 2, 3, 4 for your current understanding of each word. If you know the word, please write a sentence using it or the meaning of the word in the blank space.

1) I do not know this word.
2) I have seen the word before, but I am not sure what it means.
3) I understand this word when I see or hear it in a sentence, but I don`t know how to use it.
4) I know this word and can use it in my speaking and writing.

Vocabulary Word	Score 1,2,3,4	Definition or sample sentence using the vocabulary word to show you can use it
accurate		
analyze		
asteroid		
controversy		
evolve		
factor		
genetic		
genome		
identical		
intellectual		
majority		
mammal		
multiply		
offspring		
pesticide		
regulate		
reinforce		
stricken		
vast		
vegetarian		