Computer Assisted Language Learning Electronic Journal (CALL-EJ), 25(3), 172-194, 2024

The Effects of Interleaved Spaced Repetition Learning on Vocabulary Knowledge

Louis Lafleur (louislafleur333@gmail.com)
Kwansei Gakuin University, Japan

Yu Kanazawa (yu.kanazawa.hmt@osaka-u.ac.jp)
The University of Osaka, Japan

Abstract

Spaced repetition in Second Language Acquisition (SLA) research is a popular area of study, but few studies explore the learning of various aspects of word knowledge (Nation, 2001). Interleaved Spaced Repetition Software (ISRS) combines the principles of task interleaving (i.e., the reoccurring practice or study of multiple skills/concepts) and spaced repetition (i.e., interval-based study). This study enrolled 74 Japanese university EFL learners over two academic semesters to assess their acquisition of the New Academic Word List's (NAWL; Browne et al., 2013) word items supplemented with contextualized sentences, word/sentence audio and L1 translations (eNAWL; Kanazawa & Lafleur, 2023) while utilizing Interleaved Spaced Repetition Software (ISRS; i.e., digital flashcard study software) developed by the author. Some important findings were the uneven word knowledge pretest scores: "Meaning" (24.34%), "Form" (20.53%) and "Use" (13.16%), and the relatively even posttest score gains as encouraged by ISRS's task interleaving algorithm: "Meaning" (+16.71%), "Form" (+15.39%), and "Use" (+13.16%). Finally, the treatment group significantly outperformed the control group in total score gains (p = .002, r = .360). These results present a reasonable argument regarding of task interleaving inclusion viability within spaced repetition systems to promote a balanced and deeper learning of vocabulary alongside longer-term retention.

***Keywords*:** CALL, interleaving, spaced repetition, word knowledge, quantitative study

Introduction

Interleaved Spaced Repetition (ISR) which is the reoccurring practice or study of multiple skills/concepts (e.g., the various aspects of word knowledge; Nation, 2001) across multiple spaced and task-themed intervals has the potential become a new study standard. Although numerous spaced repetition studies have observed the learning effect of singular task-themed study items/flashcards (Kim & Webb, 2022; e.g., L2 to L1 word meaning recognition or recall), few studies have tested the various aspects of word knowledge proficiency of language learners (e.g., form/use recall); moreover, to the knowledge of the authors, none have studied the specific learning outcomes of ISR.

The purpose of this research paper is twofold. The first purpose is to review key research-informed pedagogical principles related to the fields of cognitive psychology such as the spacing effect, spaced repetition algorithms, task interleaving, and Second Language Acquisition (SLA) such as the various aspects of word knowledge, vocabulary breadth and depth to inform the development of Interleaved Spaced Repetition Software (ISRS; section 3.4 & Figure 3) specificially for the purposes of vocabulary acquisition. The second purpose is to test the viability of ISRS regarding the acquisition of various aspects of word knowledge (i.e., meaning, form, and use; Nation, 2001) in the area of Second Language Vocabulary Acquisition (SLVA; Elgort & Nation, 2010).

Literature Review

Cognitive Psychology

The Spacing Effect and the Forgetting Curve

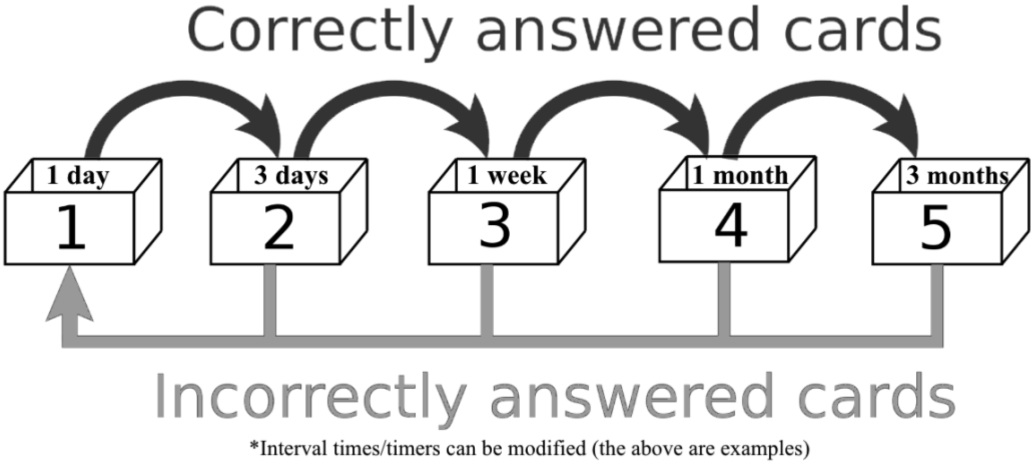
Hermann Ebbinghaus (1885/1964) established the ground-breaking “spacing effect” and “forgetting curve” theorems. The former accounts for why learners have better memory retention when they engage in spaced learning (i.e., multiple but short study sessions) compared to when they engage in massed learning (i.e., a single or few long study sessions). Modern neuroscience has confirmed that spacing repetitions at appropriate intervals allow for enough time for neuro-chemical regeneration which is necessary for establishing and strengthening brain connections (Baddeley, 1990). The latter “forgetting curve” demonstrates the exponential loss in retention strength that occurs right after learning or reviewing before slowly stabilizing into a slower and slower decline with time (Brink, 2008). For a learner, this means that newly learned information needs to be reviewed on shorter/achievable intervals at first, before being reviewed on longer and longer intervals as memory strength increases. This type of study is commonly known as spaced repetition and usually implemented with either paper or digital-based flashcards.

Spaced Repetition and the Leitner system

Flashcard-based spaced repetition is a method of study where the learner is faced with a question or a prompt (flashcard side A) for which they must try to recall or guess an appropriate answer before confirming it (flashcard side B). In terms of language acquisition, Fitzpatrick, Al-Qarni, & Meara’s (2008) study showed that such flashcard learning permitted learners to memorize a large number of words in a short time. However, Nation (2001) noted that the effectiveness of flashcard learning ultimately depends on the implementation, or lack thereof, of learning strategies by the user. Sebastian Leitner (1972) systemized sound learning strategies into a spaced-interval-based box/compartment flashcard study system capable of organizing flashcards across multiple intervals while allowing at one’s convenience the addition of new items into the mix. Specifically, the Leitner system takes advantage of the spacing effect by utilizing a “learning box” with five compartments (each with its own scheduled review timer, e.g., the 3rd box’s flashcards should be reviewed once a week) which enables word/study flashcards to move up a compartment when successfully reviewed or go back to the first box/compartment when unsuccessfully reviewed (see Figure 1).

Figure 1

The Leitner System 1972 (adapted from Zirguezi, 2012)


Correctly answered cards
1 day
1
3 days
2
1 week
3
1 month
4
3 months
5
Incorrectly answered cards
*Interval times/timers can be modified (the above are examples)

This type of system enables learners to efficiently devote more of their time to studying flashcards/items that require more review sessions to be memorized (i.e., more repetitions to be acquired in their longer-term memory) and less on the flashcards/items that do not. The Leitner system has remained popular to this day, Godwin-Jones (2010) observed that Leitner's flashcard learning box spaced repetition system became the basis or inspiration for spaced repetition software (e.g., Anki). The only undisclosed consideration of the Leitner system is the task of setting/choosing amongst the various spacing algorithms (e.g., equal, expanding) and setting the specific interval times which are left to the individuals to decide.

Massed, Equal, and Expanding Spacing

In spaced repetition literature, in terms of study/retention efficiency, there is agreement that spaced learning is more efficient than massed learning or cramming (Schütze, 2017). However, there are two different schools of thought about how to spread out study intervals within spaced learning, expanding spacing (i.e, ever increasing spacing between intervals) or equal spacing (i.e., equal/uniform spacing between intervals; see Table 1).

Table 1

Expanding, equal, and massed algorithms (adapted from Lafleur, 2020)

Algorithm type	Initial study	Interval 1	Interval 2	Interval 3	Interval 4
Expanding (x type) (~12h start → x 2)	day 1 (start point)	day 1 or 2 (~12 hours)	day 3 (1 day)	day 5 (2 days)	day 9 (4 days)
Equal (uniform) (→ every 2 days)	day 1 (start point)	day 3 (2 days)	day 5 (2 days)	day 7 (2 days)	day 9 (4 days)
Massed learning (cramming)	Total study time compressed into a single session E.g., If a study session lasts 5 minutes: 5 consecutive sessions x 5 = 25 minutes total				

In Balota, Duchek & Logan’s (2007) critical review study of massed, equal, and expanding study groups, an immediate posttest following the experiment revealed the massed pattern yielded the lowest learning efficiency, and no statistically significant difference was found between equal and expanding-spacing patterns. In contrast, Schuetze and Weimer-Stuckmann’s (2010; 2011) comparative study showed the equal-spacing pattern outperforming the expanding-spacing pattern at 83% retention versus 59% in a nine-month-delayed posttest.

However, contrary to common assumptions, Nakata's (2015) study found that equal spacing did not consistently outperform expanding spacing. Specifically, when posttest scores were assessed with some leniency, expanding spacing showed a limited but statistically significant advantage over equal spacing, resulting in a 4.6% score increase (p = .026, Np2 = .05, d = 0.12). This advantage diminished when scored strictly, with no significant difference observed (p = 0.52, Np2 = .05). Moreover, the advantage of expanding spacing over equal was also demonstrated when contrasting immediate to delayed posttest results (p = .038 for sensitive scoring, p = .044 for strict scoring) in terms of receptive knowledge scores. It should also be noted that in contrast to Schuetze & Weimer-Stuckmann (2010; 2011), Nakata (2015) controlled for absolute spacing equality between the expanding and equal conditions, ensuring an equal amount of elapsed time between the first and final intervals for both conditions.

Kim & Webb (2022) in their meta-analysis on the effects of spacing experiments SLA (i.e., 98 effect sizes collected from 48 experiments; n= 3411) echo similar results to the previously mentioned studies: (1) spacing improves second language learning on a medium to large effect size scale; (2) although shorter spacing is as effective as longer spacing on immediate posttest, it was not as effective as longer spacing in delayed posttests; (3) equal and expanded spacing results were statistically similar; and (4) variability across the various studies included in the meta-analysis could be explained by differences in their methodological implementations (e.g., number of learning sessions, feedback timing, and so on).

Moreover, on the subject of expanding versus equal spacing, Nakata (2020) stated that expanding spacing may also positively affect learners’ motivation as it may lead to higher recall success than equal spacing during the learning phase (i.e., the first intervals). In fact, there are three types of expanding algorithms: (1) + type algorithm entails spacing to increase by a factor of an addition between intervals; (2) x type algorithm entails spacing to increase by a factor of a multiplication between intervals; and (3) b^n type entails spacing to increase by a factor of exponentiation between intervals. Lafleur (2020) noted that expanding spacing algorithms +, x, and b^n are more practical in handling numerous flashcards as these can be pushed further back more aggressively in later intervals (→ monthly, etc.) as this helps to alleviate the total review burden to allow new flashcards to be introduced into the study mix more easily. That being said, there is still no clear optimal learning algorithm, and thus more research in this area is required. Lafleur (2020) also suggested that future research should perhaps focus on comparing algorithms that have not been compared much such as the various expanding algorithms amongst themselves (e.g., +, x, and b^n types; see Table 2) which could potentially lead to interesting results and inform spaced learning practices.

Table 2

+, x and b^n expanding algorithm examples (adapted from Lafleur, 2020)

Algorithm type	Initial study	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5
Expanding "+ type" (previous# + 2 days)	day 1 (start point)	day 3 (2 days)	day 7 (4 days)	day 13 (6 days)	day 21 (8 days)	day 31 (10 days)
Expanding "x type" (x 2)	day 1 (start point)	day 1 or 2 (~12 hours)	day 3 (2 days)	day 7 (4 days)	day 15 (8 days)	day 31 (16 days)
Expanding "b^n type" (E.g. ~19sec$^{(\#)}$)	day 1 (start point)	day 1 (19 secs)	day 1 (6 mins)	day 1 (2 hours)	day 3 (36 hours)	~day 31 (28½days)

Task Interleaving

Task interleaving can be described as the reoccurring practice of multiple skills or concepts. According to Nakata & Suzuki (2019), research in cognitive psychology has shown that interleaving facilitates learning better than blocking which is the practice of one skill at a time. In this same study which enrolled 115 Japanese learners of English studying five grammatical structures, similar results to previous non-L2 learning specific studies were obtained. Interleaved practice led to higher scores than blocked practice on a delayed 1-week grammatical judgment test despite the fact it led to a higher number of incorrect responses given during the treatment/practice phase. Another interesting aspect of their study was the inclusion of a third learning condition called increasing or increased practice. It is a hybrid approach between blocking and interleaving where the first sessions utilize blocking but the final sessions utilize interleaving (see Table 3). Increased practice produced results that were not statistically different from the interleaving condition, but may potentially hold future promise as its treatment/practice duration time was shorter (M = 18.16 minutes) in contrast to the interleaving's (M = 20.26 minutes) in the study.

Table 3

Task practice schedule example: blocking, increasing, and interleaving

Type	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12
Blocking	A	A	A	A	B	B	B	B	C	C	C	C
Increasing	A	A	B	B	C	C	A	B	C	A	B	C
Interleaving	A	B	C	A	B	C	A	B	C	A	B	C

Note. Verb tense practice example: A = Simple past, B = Past progressive, C = Present perfect

The concepts of interleaving and increasing can be adapted to the teaching and learning of multiple aspects of vocabulary knowledge.

Second Language Vocabulary Acquisition

The 3 Tiers (Meaning, Form, and Use) and 9 Aspects of Word Knowledge

Nation (2001) stresses the importance of considering the three tiers of word knowledge (Meaning, Form, and Function/Use) and the subsequent nine aspects of word knowledge which can be subsequently broken into receptive and productive areas of mastery for a more balanced approach in teaching and learning vocabulary (see Figure 5).

Schmitt (2008) implied that different teaching approaches might be more beneficial for different stages of word (lexical feature) knowledge acquisition as some are mastered earlier than others; therefore, focusing on the form-meaning link at first and later enhancing context(use) may prove to be effective. Both Schmitt and Nation stressed the importance of considering all three tiers of word knowledge (meaning, form, and use) when teaching and learning vocabulary.

Vocabulary Breadth and Depth

Some learners/educators may choose to place less importance on vocabulary study depth (i.e., less effort in the study of various aspects of word knowledge for each word) in order to place more importance on vocabulary breadth (i.e., more effort in the study of many words at a basic level). The former entails a slow yet deep study of vocabulary (i.e., more time spent on every single word), while the latter entails a quick yet superficial study of vocabulary (i.e., less time spent on every single word). A typical example of vocabulary breadth study would be the unilateral use of a quick receptive knowledge flashcard recall exercise (side A = L2 word, and side B = corresponding L1 word). On the other hand, vocabulary depth study typically includes elements in addition to the corresponding L1/L2 words such as L1/L2 definitions, L1/L2 example sentences, audio, and images/video. These additional elements can enable a wider range of word knowledge study, such as focus on forms, four language skills coverage, both receptive and productive knowledge, and different task flows (e.g., L1 to L2, L2 to L1, and L2 to L2). The addition of all these elements in a study system can amount to a considerable quantity of work for educators. Perhaps the most time-consuming is the addition of example sentences because of various considerations the decision process may involve, such as creating learner-level appropriate impactful sentences and providing translations.

That is not to say that both vocabulary study breadth and depth cannot be combined. In fact, this study is a prime example of such a combination as multiple word items (i.e., 963 word items) and tasks (i.e., 6 interleaved tasks; see Figure 4) were implemented as an effective way to allow for both breadth and depth of vocabulary study. However, such an approach resulted in significantly increasing the total study period to cover the material (i.e., NAWL word items divided into two academic semesters x 12 weeks, for a total of 24 weeks).

Rationale for the Current Study

Spaced repetition in Second Language Acquisition (SLA) research is a popular area of study, but few studies explore the learning of various aspects of word knowledge: “Meaning”, "Form", and "Use" (Nation, 2001; see Figure 4). To address this gap, the main author of this study developed Interleaved Spaced Repetition Software (ISRS) which combines the principles of task interleaving (i.e., the reoccurring practice of multiple skills or in this case the learning of various aspects of word knowledge) and spaced repetition/learning (i.e., interval-based study). Considering this is the first study testing the effect of ISRS on word knowledge acquisition, it was decided to compare a treatment group of ISRS to a traditional control group (i.e., without treatment) to assess its larger impact and viability in the SLA context. Moreover, in consideration of the various novel aspects and unknowns of this study, the following research questions were preferred to hypotheses to assess the learning outcomes of ISRS:

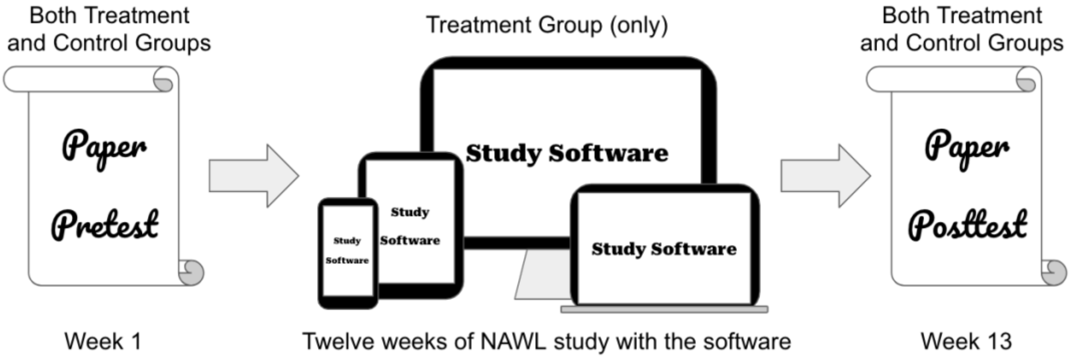
1. What aspects of word knowledge are more likely to be known before the ISRS study?
2. What aspects of word knowledge are acquired with/without ISRS?

Methodology

This study used a quantitative study approach in regard to data collection (e.g., software treatment task results and pre/posttest scores). The overall study followed the following pattern twice as the study of New Academic Word List (NAWL) word items was divided across two semesters (see Figure 2):

Figure 2

Study flow with the timeline



Context and Participants

The author recruited 74 CEFR A2/B1 level Japanese learners of English majoring in business administration at a private Japanese University. On average, participants were enrolled in four classes of either English for General Purposes (EGP) or English for Specific/Academic Purposes (ESP/EAP) with a focus on Business English. The data for this study was collected over two academic semesters (i.e., fall and spring) from a total of 74 participants who completed both pre/posttests (i.e., treatment n=64, control n=10). The treatment group completed the minimum requirement of 1000 New Academic Word List (NAWL; Browne et al., 2013) word items study tasks with the study software over a period of twelve weeks. Regarding ethical considerations, the control group during the treatment phase was given different study materials (i.e., formulaic expression learning software; e.g., “on the other hand”, “for instance”, “and so on”) to ensure comparable yet different learning opportunities. Unequal group sizes and proficiency levels (i.e., pre-test scores) were due to a convenience sampling approach being implemented (i.e., quasi-experimental study) to assure that participants from the same classroom had access to the same study condition (i.e., treatment or control) to avoid confusion among participants.

The Wordlist

The reasons behind the selection of the NAWL (i.e., 963 word items; Browne et al., 2013) for this study were fourfold: (1) the relatively high difficulty of NAWL word items which were mostly unknown among similar ESL/EFL participants in a previous study from Kanazawa & Lafleur (2023) to help avoid data/result ceiling effects; (2) a five-hundred-word study limit per semester or in this case a thousand words over two semesters was set for this research considering both participant study loads and also the time required to supplement NAWL items with example sentences and L1 translations (see Supplementary Materials 3); (3) the NAWL utilizes the flemma Word Counting Unit (WCU) which presents/counts derivational forms of words as separate entries and is considered as being a more appropriated WCU for ESL/EFL learners (e.g., McLean, 2018) than the Word Family WCU which does not include derivational word forms as additional entries and is used by lists such as the Academic Word List (AWL; Coxhead, 2000); (4) in order to infer total word gains accurately between pre/posttests, a logically sequenced wordlist (i.e., one that provides item sequencing data such as item frequency and/or difficulty) from which representative test items can be selected was required.

The previously published NAWL frequency data and prior results from Kanazawa & Lafleur (2023) permitted a hybrid re-sequencing of the word items based on both concepts (i.e., word frequency & difficulty). First, the items were sequenced into frequency-based word bands (i.e., subsets) of 100 word items according to Browne et al.’s (2013) results. Second, the items within each band/subset were re-sequenced withing their band from easiest to most difficult according to the yes-no receptive knowledge data results from 222 Japanese university students (Kanazawa & Lafleur, 2023). Moreover, word items were supplemented with translations and contextualized sentences for both testing and pedagogical applications.

The resulting updated wordlist, named eNAWL by Kanazawa & Lafleur, was utilized for both the testing and study treatment sections of this study. Half of the word bands were studied by participants in the fall semester (items 1~500), and the other half were studied by participants in the spring semester (items 501~963).

The Pre/Posttest (Pre/Post Repeated Paper Test)

Pre-recorded aurally driven repeated pre/posttests were conducted immediately before and after each twelve-week software treatment period. The number of items evaluated in these 25-minute paper tests followed McLean, Stewart, & Batty’s (2020) recommendation to test at least 40 per 1000 word items when assessing L2 meaning-recall and/or L2 form-recall modalities as this was sufficient in their bootstrapping study to reach a Cronbach’s alpha value of .90. This equated to testing 20 word items of the 500 studied in the fall semester (see Supplementary Materials 1), and 19 word items of the 463 studied in the spring semester (see Supplementary Materials 2). In an effort to balance test difficulty, the 13th, 38th, 63rd, and 88th difficulty-ranked word items of each 100-word band were selected to comprise the test with the exception of the 88th word item of the final band since it only comprised a total of 63 words. These tests were comprised of three sections for each of the selected word items (see Table 4 and Supplementary Materials 1 & 2):

Table 4
Pre/posttest sections and task flow

Test sections	Task flow	Similar Test Format
(1) "Meaning" Listening Recall	After listening to a sentence and target word audio in English, the participant was asked to translate/write the target word in Japanese.	(McLean et al., 2021) Spoken Receptive Meaning-Recall /Listening meaning-recall
(2) "Form" Dictation Recall	After listening to the target word audio in English, the participant was asked to write its basic/dictionary form in English.	(Cheng & Matthews, 2018) Testing productive/phonological (ProPhon) vocabulary knowledge
(3) "Use" Listening Recall	After listening to the sentence audio in English of the target word, the participant was asked to write its translation in Japanese.	None, but inspired by Nation's (2001) suggestion to enable a more "in-depth" learning/testing of vocabulary to ensure the correct "use" of the words.

In an effort to uphold fair results, the grader (i.e., main author) was blind to which of the groups (i.e., treatment, and control) each of the tests was associated with. The tests were graded using the following logic: incorrect responses were attributed a score of zero, debatable/arguable responses were attributed a score of 0.5 (e.g., a slight omission or mistake that did not overtly compromise an answer in test section 1 or 3), and correct responses were attributed a score of one. Moreover, part-of-speech variations in answers, and alternate viable answers (e.g., synonyms) were accepted when deemed appropriate. In contrast, the most strictly graded was section 2 as English spelling mistakes in answers were attributed a score of zero.

The Software

Numerous studies such as Suzuki (2019), Li (2015), and Vatz et al. (2013) point to the possible optimization of SLA with the implementation of individualized learning approaches for learners as one type of instruction can be more or less effective than the next depending on individual learner weaknesses and strengths.

Interleaved Spaced Repetition (ISR), which combines the concepts of spaced repetition (i.e., interval-based study) and task interleaving (i.e., the reoccurring practice of multiple skills or concepts), also follows this trend to further individualize study. The main author's ISR permits learners to not only devote more time/attention to the more difficult (word) items but specifically to the more difficult aspect(s) of each specific (word) item. In other words, the interleaving algorithm customizes review/study for each user's specific learning needs (i.e., users who struggle with the spelling of a particular word item will encounter this specific spelling task more often, and users who do not struggle with it will not encounter it as often) (Lafleur, 2020; see Figure 3).

Figure 3

Interleaved Spaced Repetition Software ISRS tasks (adapted from Lafleur, 2020)

Tier Q# level	Flow L1= native language L2= second language	Task Type	Task/Interval Route: ↓ when answered successfully ⟲ or ←when answered unsuccessfully (cooldown time/next review; m= minute, h= hour, d= day)		
Meaning Q#1 word or phrase	L2 audio to L1 recall	Recall Check	Session 1 (Q#1) ⟲(start/⟲=5m) ↓	Session 7 (Q#1) ← (6d) ↓	Session 13 (Q#1) ← (162d) ↓
Meaning Q#2 word or phrase	L1 word/phrase to L2 recall	Recall Check	Session 2 (Q#2) ⟲ (8h) ↓	Session 8 (Q#2) ← (9d) ↓	Session 14 (Q#2) ← (243d) ↓
Form Q#3 word or phrase	L2 audio to L2 word/phrase	Spelling	Session 3 (Q#3) ⟲ (16h) ↓	Session 9 (Q#3) ← (18d) ↓	Session 15 (Q#3) ← (486d) ↓
Form Q#4 sentence	L2 (blank) to L2 sentence	Fill the blank	Session 4 (Q#4) ⟲ (1d) ↓	Session 10 (Q#4) ← (27d) ↓	Session 16 (Q#4) ← (729d) ↓
Use Q#5 sentence	L2 sentence to L1 sentence	Writing	Session 5 (Q#5) ⟲ (2d) ↓	Session 11 (Q#5) ← (54d) ↓	Session 17 (Q#5) ← (1458d) ↓
Use Q#6 sentence	L1 sentence to L2 sentence	Writing	Session 6 (Q#6) ⟲ (3d) ↓	Session 12 (Q#6) ← (81d) ↓	Session 18 (Q#6) ← (2187d) end
(Optional)* Q#7 Text	L2 Listening and L2 Reading	(Voiced) Reading	↳ back to top ↑ *optional, completing a set of words could trigger Q#7	↳ back to top ↑	

Note. Only one task/question type (Q#) is shown/asked with each study/review session; after a successful answer follow the ↓ arrow or ⟲ / ← arrow after an unsuccessful answer; see Figure 4 for question/task details.

Each word card item follows the task/interval route independently according to its most recent successful/unsuccessful response which decides its following session number, study/review cooldown, and task/question. During a user's first-time use of ISRS, they will only encounter Q#1 type questions for word items (i.e., unless they use the software for more than eight hours straight and some of their word items reach session#2). However, as they space their study (i.e., take study breaks), they will encounter other question types as cooldown timers for studied word items continuously run no matter if users are logged on or off. Moreover, the author's ISRS system was programmed to fully prioritize review (i.e., word cards that have run down their study/review cooldown timer and are again ready for study) before adding new word card items into a user's ISRS study mix (i.e., introducing new items for the first time; initial session 1).

Another important consideration was setting an appropriate review timer in the case of incorrect responses in ISRS's study sessions #1 and #7 (see Figure 3) as these are the only sessions that are likely to be reviewed within the same study session. It was decided to set a relatively long spacing timer (i.e., 5 minutes) between incorrect responses and review. Results from Nakata, Suzuki & He (2022), which reported on within-session spacing, support a longer spacing approach as it was found that despite a slight increase in treatment duration (i.e., a higher number of attempts required before reaching a successful review, which was controlled/accounted for by implementing an ANCOVA calculation), long spacing led to higher scores and better long-term retention for both initial and relearning intervals on posttests.

For ISRS, the implementation of an x-type expanding spacing algorithm (see Table 2 & Figure 3) was inspired by a combination of Nakata's (2015) research results and the fact that expanding algorithms are more practical in handling multiple study items at once (e.g., ~100 and more) as these can be pushed back more aggressively in later intervals (e.g., monthly, and so on.) which permits new items to be added more easily (Lafleur, 2020).

Cooldown times need to be carefully set for ISRS intervals. Beta-versions of the software were tested with two different approaches (i.e., shorter and longer interval cooldown times). Both versions followed the same expanding algorithm and interval spacing was increased by a factor of a multiplication (i.e., an alternating factor of x1.5 or x2 from interval to interval). The former "shorter" version had an initial cooldown time of 4 hours for session#2 which resulted in much shorter interval cooldown times for subsequent intervals (e.g., session#3 = 8 hours, session#4 = 12 hours, session#5 = 1 day, and so on). The former was found to create a study system which focused slightly too heavily on review and not enough on integrating new cards/word items into the study mix. In other words, it was found that test users would only encounter about 300 word items during a study semester, whilst test users using the latter "longer" current version would on average encounter almost all 500 word items at least once during a study semester.

Specifically, in terms of vocabulary teaching and learning, the interleaving component of ISR was inspired by Paul Nation's aspects of word knowledge (see Figure 5). For this research project, it was decided to include six interleaved study tasks (i.e., two tasks per tier of word knowledge, meaning, form, and use; see Figures 3 & 4) as this was judged to cover a major portion of Nation's aspects of word knowledge (see Figure 5).

Participants could freely access the study's wordlist and software through a free web-based learning application accessible 24/7 via any web browser at www.eigomemo.com for any type of device (e.g., smartphones, computers, and so on; i.e., responsive website application design). Participants were recommended a weekly goal of ~200 study tasks a week and a 12-week mark aspirational goal of 2400 tasks (i.e., completing one study task involves responding to one question related to one eNAWL word item as determined by the software's algorithm).

Figure 4

Interleaved Spaced Repetition (ISR) task/question details

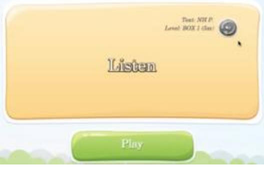
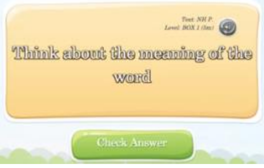
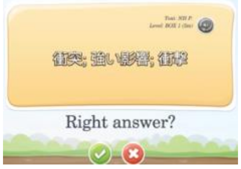
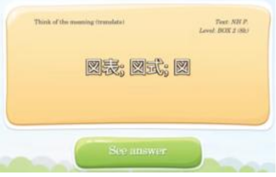
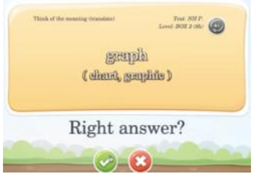
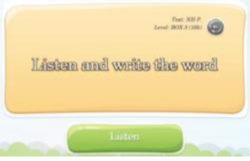
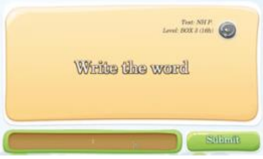
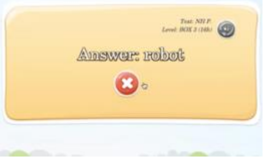
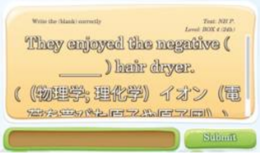
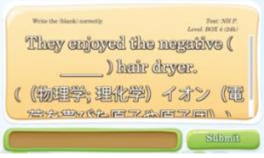
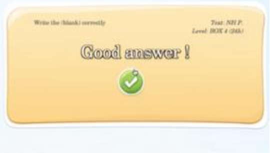
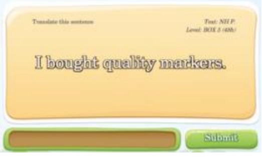
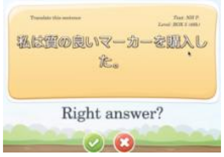
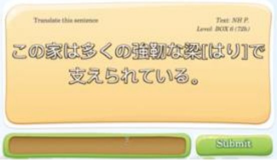
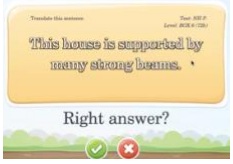
Q#1 e.g., (word) impact	 Listen Play 1. “Listen” is displayed. 2. The user must push on the (play) button.  Think about the meaning of the word Check Answer 3. “Think about the meaning of the word” is displayed. 4. The sentence/word audio is played. 5. The user thinks and when ready clicks on (check answer).  衝突; 強い影響; 衝撃 Right answer? 6. Recommended answers are displayed. 7. The user self-assesses the validity of their answer/recall. 8. The user chooses right or wrong (honor system).
Q#2 e.g., (word) graph	 Think of the meaning (translate) 図表; 図式; 図 See answer 1. “Think of the meaning (translate)” is displayed. 2. The L1 word(s)/synonyms are displayed. 3. The user thinks about a valid corresponding L2 word, and clicks on (see answers)  graph (chart, graphic) Right answer? 4. Both recommended and viable answers are displayed. 5. The user self-assesses the validity of their answer/recall 6. The user validates or refutes their answer (honor system)
Q#3 e.g., (word) robot	 Listen and write the word Listen 1. “Listen and Write the word” is displayed. 2. The user must push on the (listen) button.  Write the word Submit 3. “Write the word” is displayed. 4. The sentence/word audio is played. 5. The user thinks and writes/spells the word and clicks (submit).  Answer: robot 6. The software automatically checks the user’s answer. 7. “Good or wrong answer” is displayed. 8. The user must click on the flashcard to move on to the next task.
Q#4 e.g., (word) ion	 Write the (blank) correctly They enjoyed the negative (_____) hair dryer. ((物理学; 理化学) イオン (電 Submit 1. “Write the blank (correctly)” is displayed. 2. The target sentence with a (blank) and L1 hints are shown.  Write the (blank) correctly They enjoyed the negative (_____) hair dryer. ((物理学; 理化学) イオン (電 Submit 3. The user thinks about the answer and writes the missing word. 4. The user then clicks on (submit).  Write the (blank) correctly Good answer ! 5. The correct answer is shown. 6. The system automatically compares their answer with viable answers. 7. A click on the screen is necessary to move on to the next task
Q#5 e.g., (word) marker	 Translate this sentence I bought quality markers. Submit 1. “Translate this sentence” is displayed. 2. The target L2 sentence is displayed. 3. The user thinks about and writes the sentence in their L1 and clicks (submit).  Translate this sentence 私は質の良いマーカーを購入した。 Right answer? 4. A recommended answer is displayed. 5. The user validates or refutes their answer (honor system).
Q#6 e.g., (word) beam	 Translate this sentence この家は多くの強靭な梁[はり]で支えられている。 Submit 1. “Translate this sentence” is displayed. 2. The target L1 sentence is displayed. 3. The user thinks about and writes the sentence in their L2 and clicks (submit).  Translate this sentence This house is supported by many strong beams. Right answer? 4. A recommended answer is displayed. 5. The user validates or refutes their answer (honor system).

Figure 5

Aspects of word knowledge covered by ISRS tasks (adapted from Nation, 2001)

Type	Sub-type		(Nation, 2001) explanation/example	ISRS task#	
Form	Spoken	R	What does the word sound like?	○	1, 2, 3, (5, 6)
		P	How is the word pronounced?	△	if 7 included
	Written	R	What does the word look like?	○	2, (3, 4, 6)
		P	How is the word written and spelled?	○	3, 4, 6
	Word parts	R	What parts are recognizable in the word?	△	if highlighted
		P	What word parts are needed to express the meaning?	×	not included
Meaning	Form & meaning	R	What meaning does this word form signal?	○	1, 5
		P	What word form can be used to express this meaning?	○	2, 4, 6
	Concepts & referents	R	What is included in the concept?	△	if included
		P	What items can the concept refer to?	△	if included
	Associations	R	What other word does this make us think of?	○	2, 4, 6
		P	What other words could we use instead of this one?	○	2, 4, 6
Use	Grammatical Functions	R	In what patterns does the word occur?	○	if covered 4, 6
		P	In what patterns must we use this word?	○	if covered 4, 6
	Collocations	R	What words or types of words occur with this one?	○	4, 6, (5)
		P	What words or types of words must we use with it?	○	6
	Constraints on use	R	Where/When/How often would we expect to meet it?	△	4, 6, (5)
		P	Where/When/How often can we use this word?	△	6

Note. R= Receptive, P= Productive; ( )= shown with wrong answer; and ISRS task coverage rate as judged by the author ○= good, △= average/possible, ×= poor

The Analysis Procedures

Skewness, Kurtosis, and Shapiro-Wilk tests were conducted in SPSS for the purpose of verifying the normality of distributions in the data. Pretest scores did not reveal any problems; however, software participation and posttest results revealed several instances of concern such as Shapiro-Wilk scores of over .05, skewness scores below -1 or over +1, and z-scores over 3 standard deviations. To address these issues, the author decided to use non-parametric calculations such as median/IQR calculations for descriptive statistics (i.e., somewhat similar yet different to mean and standard deviation calculations as they control better for outliers in the data), and Mann-Whitney *U* tests for statistical significance testing to circumvent abnormally distributed data. Moreover, pre/posttest gain scores were preferred to pre/posttest results when calculating the effect of the software in an effort to control for varying participant proficiency levels.

For non-parametric data effect size calculations, the author followed Fritz et al.'s (2012) recommendation to follow Cohen's (1988) equation that utilizes the *z* value (i.e., Mann-Whitney *U's* "standardized test statistic"): $r = z / \sqrt{N}$ (*r* effect size = *z* value divided by the square root of the sample number).

Results

The collected data from the treatment software and pre/posttests were used to create the following tables in an effort to reveal study habits and word knowledge acquisition.

Table 5 "Average within-semester software study participation" shows the participants' median results in terms of participation totals in flashcards/tasks studied. The IQR, min, and max numbers for all data points show there was a substantial amount of variation between the participants, which in all probability contributed to the study's abnormally distributed data.

Table 5

Average within-semester software study participation (n= 64)

Data point	Median	IQR	Min	Max
#tasks/questions studied	2377.00	288.00	1043	3005
#active study days	22.50	18.50	6	77
#tasks/questions per active day	91.15	75.13	33.42	258.17
#total study minutes	611.00	244.50	223	1344
#tasks/questions per minute	3.71	1.20	1.79	5.14

Table 6 "Responses per ISRS question type" shows the number of responses and success ratio for every individual task and aspect of word knowledge type. Overall, the success ratio was generally low (43.48% overall) as these word items were taken from a frequency-based academic wordlist, and many word items were unknown to the participants (see pretest results; Table 12). The question types with the lowest success were productive L2 written form(s) input tasks (i.e., spelling 29.88% and fill-the-blank 27.68%).

Table 6

Responses per ISRS question/task type (n= 64)

ISRS ?type	[Question Type] Question Flow	# of Tasks Completed	% of Tasks Completed	Successful Recall#	Success Rate %	Success Rank /6
Q1	[Recall Check] L2 audio to L1 recall	45070	29.41%	22171	49.19%	3rd
Q2	[Recall Check] L1 word to L2 recall	32871	21.45%	18839	57.31%	2nd
Q3	[Spelling] L2 audio to L2 word	38328	25.01%	11451	29.88%	5th
Q4	[Fill-the-blank] L2 blank to L2 sentence	22495	14.68%	6226	27.68%	6th
Q5	[Writing] L2 sentence to L1 sentence	7744	5.05%	4741	61.22%	1st
Q6	[Writing] L1 sentence → L2 sentence	6728	4.39%	3197	47.52%	4th
	TOTAL	153236	100.00%	66625	43.48%	Overall

Table 7 “Responses per ISRS Tier/Main Aspect of Word Knowledge” shows the number of response totals and success ratio averages for tasks related to the same tier/main aspect of word knowledge. Lower success scores under form(s) are likely due to tasks Q#3 & Q#4 being evaluated automatically and strictly (i.e., no partial credit was given) by the software.

Table 7

Responses per ISRS tier/main aspect of word knowledge (n= 64)

ISRS task type	Tier/Main Aspect of Word Knowledge Focus	# of Total Responses	% of Total Responses	Successful Recall#	Success rate %	Success Rank /3
Q1 & Q2	Meaning	77941	50.86%	41010	52.62%	2nd
Q3 & Q4	Form(s)	60823	39.69%	17677	29.06%	3rd
Q5 & Q6	Use	14472	9.44%	7938	54.85%	1st
TOTAL		153236	100.00%	66625	43.48%	Overall

Table 8 “Responses per ISRS memory tiers” (i.e., short, mid and long-term as defined by the main author) shows the number of responses and success ratio for each interval box/session regarding eNAWL word items. An important factor is the higher success rate when tasks were repeated in the next tier (i.e., 41.68% to 91.70%). It should be noted that no responses were collected for interval times between 162 ~ 2187 days as the study treatment for individual NAWL word items was limited to a period of twelve study weeks.

Table 8

ISRS consolidated memory tier results (n= 64)

Memory Tier	Sessions / Boxes	Study/Review Cooldown Intervals	# of Total Tasks	% of Total Tasks	#Correct Responses	Success %
Short-term	#1 ~ #6	initial ~ 3 days	147732	96.41%	61578	41.68%
Mid-term	#7 ~ #12	6 ~ 81 days	5504	3.59%	5047	91.70%
Long-term	#13 ~ #18	162 ~ 2187 days	0	0.00%	0	0%
TOTAL	#1 ~ #18	initial ~ 2187 days	153236	100.00%	66625	43.48%

Note. Refer to Figure 3 for more detailed look at session review cooldown timers.

Table 9 “Responses per specific ISRS Interval Box” shows the number of responses and success ratio for each interval box/session. An important factor is the higher success rate when tasks are repeated 6 intervals later from a short to mid-term memory tier interval (e.g., Question Type#1 from Box#1 to Box#7, 46.35% to 97.56%). Another important observation is the most difficult task types in the first tier (boxes #1~6), remained more or less most difficult within the second tier (boxes #7~12).

Table 9

Responses per specific ISRS interval box (n= 64)

Interval Box	Task Type	Study Interval Timer	# of Total Tasks completed	% of Total Tasks	#Tasks Rank / 10	#Correct Responses	Success %	Success Rank/ 10
#1	Q#1	Start	42569	27.78%	1st	19731	46.35%	8th
#2	Q#2	8 hours	30999	20.23%	3rd	17002	54.85%	6th
#3	Q#3	16 hours	37293	24.34%	2nd	10741	28.80%	9th
#4	Q#4	24 hours	22399	14.62%	4th	6166	27.53%	10th
#5	Q#5	2 days	7744	5.05%	5th	4741	61.22%	5th
#6	Q#6	3 days	6728	4.39%	6th	3197	47.52%	7th
#7	Q#1	6 days	2501	1.63%	7th	2440	97.56%	2nd
#8	Q#2	9 days	1872	1.22%	8th	1837	98.13%	1st
#9	Q#3	18 days	1035	0.68%	9th	710	68.60%	3rd
#10	Q#4	27 days	96	0.06%	10th	60	62.50%	4th
#11~18	Various	81 ~ 2187 days	No data	0.00%	N/A	0	0%	N/A
TOTAL	Q#1~6	start ~ 2187 days	714988	100.00%	Overall	315174	44.08%	Overall

Table 10 "ISRS eNAWL word study effect according to Pre/Posttest Score Results" shows both similarities and differences between the control and treatment groups. Both groups scored much lower on the productive "use" recall sections of their pretest than on the "meaning" and "form" recall sections. In terms of score gains, although "use" scores remained the lowest for both groups, the treatment group's "use" score had the highest relative increase (+100%). This important score increase is most likely linked to the study's software used by the treatment group which focused on the more difficult areas/aspects of word knowledge for this group's participants. Finally, the treatment group significantly outperformed the control group in total score gains (p = .002, r = .360). In terms of vocabulary gains, it can be estimated that the control group increased their academic vocabulary by 48 words and the treatment group by 68 academic words for the treatment group over the twelve-week study span. Although the vocabulary gain results at first glance do not seem too far apart, it should be noted that the relative score increase was much higher for the treatment group (+75.64%) in comparison to the control group (+39.64%).

Table 10

ISRS eNAWL word study effect according to pre/posttest score results

Group n =	Test	**“Meaning”** test score Median % (IQR %)	**“Form”** test score Median % (IQR %)	**“Use”** test score Median % (IQR %)	**Total** test score Median % (IQR %)
Group 0	Pretest	28.75% (25.00%)	32.50% (25.00%)	13.75% (29.38%)	24.17% (30.62%)
n=10	Posttest	35.00% (30.00%)	42.50% (21.25%)	21.25% (16.15%)	33.75% (26.67%)
		+6.25% (+5.00%)	+10.00% (-3.75%)	+7.50% (-13.23%)	+9.58% (-3.95%)
		+21.74% relative.diff.	+30.77% relative.diff.	+54.55% relative.diff.	+39.64% relative.diff.
Group 1	Pretest	24.34% (19.21%)	20.53% (21.05%)	13.16% (21.78%)	17.98% (19.82%)
n=64	Posttest	41.05% (24.61%)	35.92% (32.96%)	26.32% (32.53%)	31.58% (29.38%)
		+16.71% (+5.40%)	+15.39% (+11.91%)	+13.16% (+10.75%)	+13.60% (+9.56%)
		+68.65% relative dif.	+74.96% relative dif.	+100.00% relative dif.	+75.64% relative dif.
Mann-Whitney		U = 543.000	U = 397.000	U = 514.500	U = 515.500
p value		p = <.001	p = .222	p = .002	p = .002
z-derived r		r = .411	r = .142	r = .359	r = .360

Note. Group 0= control group; Group 1= treatment group.

Discussion

RQ1: What aspects of word knowledge are more likely to be known before the ISRS study?

As the data analysis shows, both the control and treatment groups scored lower on the “use” listening recall sections (13.75% and 13.16%) than on the “meaning” listening recall (28.75% and 24.34%) and “form” dictation recall (32.50% and 20.53%) sections of the pretest (see Table 10). These results are not surprising as Schmitt (2008) implied that different lexical features are mastered earlier than others. Both Schmitt (2008) and Nation (2001) stressed the importance of considering all three tiers of word knowledge (meaning, form, and use) when teaching/learning vocabulary. Their recommendations formed the basis for the development of ISRS (see Figure 3) in an effort to direct/focus vocabulary not only on the more difficult words but on the more difficult areas of word knowledge.

RQ2: What aspects of word knowledge are acquired with/without ISRS?

The results showed that the treatment group which used the software significantly outperformed the control group in total score gains (p = .002, r = .360). Both “use” listening recall (p = .002) and “meaning” listening recall (p = <.001) gain scores were significantly better. However, “form” dictation recall scores although better (i.e., +15.39% treatment vs +10.00% control) were not significantly better (p = .222). One reason for this could be the fact that the control group had practiced and focused on improving their general spelling ability while studying a formulaic word list during the treatment phase. As for other score increases such as in the control group’s “meaning” listening recall (+6.25%) and “use” listening recall (+7.50%) scores, it could be hypothesized that these were due to either a kind of test effect (i.e., test taking itself led to learning) and/or the general improvement of their English ability due their other classes/studies during the twelve-week gap between the pre/posttests. Another interesting point was the relatively even posttest score gains as encouraged by ISRS’s task interleaving algorithm: “Meaning” (+16.71%), “Form” (+15.39%), and “Use” (+13.16%) in contrast to the control group’s score gains which as previously mentioned were higher in terms of “form” dictation recall gains than in the other two categories. These results could not be compared to previous studies considering the novel aspect of this study.

Conclusion

Although there have been numerous second language acquisition studies that have studied the effect of task interleaving and spaced learning separately, this study is unique in the fact that it evaluated the effect of both concepts combined. An important finding in this study was the unevenness in the participants' grasp of the various aspects of word knowledge shown in the pretest (i.e., lower "use" scores). These knowledge gaps were addressed by the treatment study software which customized each participant's study to focus on the weaker aspects of their vocabulary knowledge. In contrast to the control group, the treatment group's lowest score category "use" had the highest relative score increase of all; +100% between pre/posttest). Overall, the treatment study participants had higher learning outcomes in terms of pre/posttest score gains ($p = .002$, $r = .360$). These results present a reasonable argument in regard to task interleaving inclusion viability within spaced repetition systems to promote a balanced and deeper learning of vocabulary alongside longer-term retention.

Limitations and Future Directions

Although this study was able to provide some insight into the research area of interleaved spaced repetition for vocabulary learning, it should be noted that it also has its share of limitations that should be addressed in future studies. Although some precautions were taken to uphold a high degree of testing scrutiny (e.g., the main author was blind to which group tests were associated while grading), it would be best to implement a multiple-test-rater approach in future studies. Second, the limited number of participants, their varying participation in using the treatment software, and the unevenness of their L2 proficiency level within both study groups contributed to data issues (e.g., normality of distribution issues in the data). These limitations were controlled for by using non-parametric means of analysis (see section 3.5). It would be best to avoid such problems in future studies from the onset by having larger and more evenly distributed participant sample sizes (i.e., in terms of both number and proficiency level). Moreover, future research should also include delayed posttests which would be more in line with recent distributed L2 vocabulary acquisition research. Finally, many avenues of research are still left unexplored, for example, future studies could compare ISRS to traditional/other spaced repetition software, compare the efficiency of various types of expanding spaced algorithms, and explore the study/learning of other study content.

References

Baddeley, A. (1990). *Human memory: Theory and practice*. Lawrence Erlbaum Associates.

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83-105). Psychology Press. https://doi.org/10.4324/9780203837672

Brink, T. L. (2008). *Psychology: A student friendly approach. Unit 7: Memory, p.126*. Retrieved August 29, 2021 from: http://www.saylor.org/site/wp-content/uploads/2011/01/TLBrink_PSYCH07.pdf

Browne, C., Culligan, B., & Phillips, J. (2013). New Academic Word List (NAWL). Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Retrieved from http://www.newgeneralservicelist.org/nawl-new-academic-word-list

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3-25. https://doi.org/10.1177/0265532216676851

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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

Ebbinghaus, H. (1964). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie* [Memory: A contribution to experimental psychology]. Dover. (Original work published 1885)

Elgort, I., & Nation, P. (2010). Vocabulary learning in a second language: Familiar answers to new questions. *Conceptualising 'learning' in applied linguistics*, 89-104. https://doi.org/10.1057/9780230289772_6

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language learning*, *61*(2), 367-413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language learning journal*, *36*(2), 239-248. https://doi.org/10.1080/09571730802390759

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). "Effect size estimates: Current use, calculations, and interpretation": Correction to Fritz et al.(2011). *Journal of Experimental Psychology: General*, *141*(1), 2-18. https://doi.org/10.1037/a0026092

Godwin-Jones, R. (2010). From memory palaces to spacing algorithms: approaches to second language vocabulary. *Language, Learning & Technology*, *14* (2), 4-11. https://doi.org/10125/44208

Kanazawa, Y., & Lafleur, L. (2023). ENAWL: Enriching the New Academic Word List with emotional valence, familiarity, and knowledgeability. *Kokusaigaku Kenkyu–Journal of International Studies*, *12*(1), 141-151. http://hdl.handle.net/10236/00030725

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269-319. https://doi.org/10.1111/lang.12479

Lafleur, L. (2020). The indirect spaced repetition concept. *Vocabulary Learning and Instruction*, *9*(2), 9-16. https://doi.org/10.7820/vli.v09.2.lafleur

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, *59*(4), 567-587. https://doi.org/10.3138/cmlr.59.4.567

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*(1), 223-250. https://doi.org/10.1075/eurosla.5.11lau

Leitner, S. (1972). *So lernt man lernen: Der weg zum erfolg* [How to learn to learn: The road to success]. Verlag Herder.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*(3), 385-408. https://doi.org/10.1093/applin/amu054

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(3), 823-845. https://doi.org/10.1093/applin/amx003

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389-411. https://doi.org/10.1177/0265532219898380

McLean, S., Raine, P., Pinchbeck, G., Huston, L., Kim, Y., Nishiyama, S., & Ueno, S. (2021). The internal consistency and accuracy of automatically scored written receptive meaning-recall data: A preliminary study. *Vocabulary Learning and Instruction, 10*(2), 64-81. https://doi.org/10.7820/vli.v10.2.mclean

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/S0008413100018260

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(1), 3-20. https://doi.org/10.1017/S0958344008000219

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*(4), 677-711. https://doi.org/10.1017/S0272263114000825

Nakata, T. (2020). Vocabulary and computer technology: A commentary on four studies for JALT Vocabulary SIG. *Vocabulary Learning and Instruction*, *9*(2), 39-47. https://doi.org/10.7820/vli.v09.2.nakata

Nakata, T., & Suzuki, Y. (2019). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *The Modern Language Journal*, *103*(3), 629-647. https://doi.org/10.1111/modl.12581

Nakata, T., Suzuki, Y., & He, X. (2022). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*, *73*(3), 799-834. https://doi.org/10.1111/lang.12553

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917-1927. https://doi.org/10.3758/BF03192925

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363. https://doi.org/10.1177/1362168808089921

Schuetze, U., & Weimer-Stuckmann, G. (2010). Virtual vocabulary: Research and learning in lexical processing. *CALICO Journal*, 27, 517-528. https://doi.org/10.11139/cj.27.3.517-528

Schuetze, U., & Weimer-Stuckmann, G. (2011). Retention in SLA lexical processing. *CALICO Journal*, 28, 460-472. https://doi.org/10.11139/cj.28.2.460-472

Schütze, U. (2017). *Language learning and the brain: Lexical processing in second language acquisition*. Cambridge University Press. https://doi.org/10.1017/cbo9781316665619

Suzuki, Y. (2019). Individualization of practice distribution in second language grammar learning: The role of metalinguistic rule rehearsal ability and working memory capacity. *Journal of Second Language Studies*, *2*(2), 169-196. https://doi.org/10.1075/jsls.18023.suz

Vatz, K., Tare, M., Jackson, S. R., & Doughty, C. J., Granena, G., & Long, M. (2013). Aptitude-treatment interaction studies in second language acquisition. *Sensitive periods, language aptitude, and ultimate L2 attainment*, *35*, 273. https://doi.org/10.1075/lllt.35.11vat

Zirguezi, 2012. Leitner system animation CC0 1.0 Public Domain, Retrieved September 23rd, 2019 from: https://en.wikipedia.org/wiki/Leitner_system#/media/File:Leitner_system_animation.gif

Acknowledgments & Ethical Statement

The authors would like to thank all those who contributed their advice and time to this paper, including the Japan Society for the Promotion of Science KAKENHI for their financial support, Grant-in-Aid for Scientific Research Grant Number JP19K00899 (https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-19K00899/)

Both authors conducted the research abiding by their institutions' ethical regulations and were granted approval by the Japan Society for the Promotion of Science KAKENHI to conduct this study. Informed consent was obtained from all individual participants involved in the study. The authors declare that they have no conflicts of interest.

About the Authors

Louis Lafleur is a lecturer at Kwansei Gakuin University conducting research related to the fields of second language vocabulary acquisition, cognitive psychology, computer-assisted language learning, and game-informed language learning (gamification).

Yu Kanazawa is an associate professor (lecturer) at The University of Osaka conducting research related to the fields of second language vocabulary acquisition, philosophy, emotion, and psychology.

Author #1 ORCiD. Louis Lafleur, https://orcid.org/0000-0003-1996-7521
Author #2 ORCiD. Yu Kanazawa, https://orcid.org/0000-0003-1398-5913

Credit Author Statement

Author#1: conceptualization (lead); data curation (lead); funding acquisition (supporting); methodology (lead); project administration (lead); software (lead); supervision (lead); vocabulary materials (equal); writing (lead); writing – review & editing (lead).

Author#2: conceptualization (supporting); funding acquisition (lead); supervision (supporting); vocabulary materials (equal); writing – review & editing (supporting).

Supplementary Materials 1:
NAWL Pre/Posttest with Answers (#1~500 range)

(Meaning)	(Form)	(Use)
例. 教科書	textbook	その歴史の教科書には多くの間違いがあった。
1.類似；類似点；（数学）相似	similarity	その双子は、お互いの類似点を楽しんでいる。
2.分子；微粒子；微量	molecule	それは分子レベルで破壊された。
3.談話；会談；講演；論説	discourse	最近の公の場の議論はおそろしい。
4.後戻り；退行；逆行；復帰；回帰	regression	幼児退行しているとして怒られた。
5.集めた；集団の；共同の；集合的な	collective	集団的自衛権が戦争を拡大してしまった。
6.有機体；有機的組織体；生命体	organism	人体（という有機体）はとてもよくできている。
7.検出；発見；発覚	detection	がんの早期発見が彼らの命を救った。
8.（大脳などの）皮質；（植物学）皮層	cortex	彼の大脳皮質は損傷していた。
9.原子（物理学，化学）；極小；微塵	atom	私は原子についてより良く知っている。
10.比較できる；匹敵する；同等な	comparable	彼女の成果は金メダルに匹敵する。
11.溶解；融合；連合	fusion	融合エネルギーを活用した兵器は人類を滅亡させることができる。
12.胎児（受胎の約３か月以降）の；	fetal	胎児の発育中に何の問題も無かった。
13.役に立つこと；利用できること	availability	その危機の間トイレットペーパーが入手不可能であった。
14.一般化；普遍化；概括；一般論	generalization	一般化によってそれは理解しやすくなった。
15.簡単に；快く；準備万端で	readily	その危機の間、トイレットペーパーがなかなか手に入らなかった。
16.注解；;論評；注釈；記録；解説	commentary	そのニュース解説[注解]はとても良かった。
17.正しく；正確に言えば	correctly	彼は正しくアルファベットを綴ることができなかった。
18.拒絶；排除；廃棄物	rejection	大統領の戦争参加拒絶は多くの命を救った。
19.腐る；虫歯にする；崩壊する	decay	最新技術が彼の歯を虫歯[朽ちること]から守った。
20.静脈；血管；植物の茎の管；気質	vein	彼女の静脈は強くて健康だ。

Supplementary Materials 2:
NAWL Pre/Posttest with Answers (#501~963 range)

(Meaning)	(Form)	(Use)
例[example]. 教科書	textbook	その歴史の教科書には多くの間違いがあった。
1.歴史的に；歴史に関して	historically	私が発見した鏡は歴史的に重要であることが分かった。
2.板挟み状態；二律背反	dilemma	板挟み状態でストレスが多くたまった。
3.洗練させる；凝る；複雑	sophisticate	残念ながら、彼女はあまり洗練されていない。
4.影響されやすい；感染しやすい	susceptible	彼女はウイルスに感染しやすい。
5.安売り；契約；安い買い物	bargain	これは損な買い物[貧乏くじ]だ。
6.持続可能な；維持できる	sustainable	持続可能な開発は私たちの未来を救うだろう。
7.矛盾する；反対の	contradictory	この良い先生は自分が言うことに反することは行なわなかった。
8.弾力のある；融通の利く	elastic	この輪ゴムは十分に弾力があってしっかりしている。
9.罰する；乱暴に扱う	punish	彼は罰せられるべき人びとを的確に特定した。
10.加工業者；処理装置，プロセッサー	processor	教授が新しい処理装置の開発に成功した。
11.余り；剰余金；黒字（金額）	surplus	私たちの会社は黒字[余剰]から赤字[不足]に転じた。
12.毛管の；毛状の；毛細血管	capillary	喫煙が彼女の毛細血管を傷つけた。
13.講義者；講師	lecturer	私たちはとても人気のある講演者[講師]を招待した。
14.社会化する；社交的にする	socialize	多くの社長と社交したため、彼は良い仕事を見つけることができた。
15.信用できること；信用性	credibility	彼女の話は信憑性が高い。
16.モグラ；スパイ；ほくろ；防波堤	mole	モグラ[スパイ]によって秘密が漏洩した。
17.巧妙な；ずるい；賢い；難しい	tricky	そのドアは一筋縄では開かないので彼女はいらいらした。
18.こする；摩擦する；する；磨く	rub	彼が目をこすったときに危険なウイルスが体に入った。
19.噴霧器；アトマイザー；煙霧質	aerosol	有害なエアロゾル[空気中の煙霧質微粒子]は完全に除去された。