*ReCALL* (2025), 37: 1, 96–113
doi:10.1017/S095834402400020X

RESEARCH ARTICLE

# Efficiency trumps aptitude: Individualizing computer-assisted second language vocabulary learning

Yuichi Suzuki
Waseda University, Japan (yszk@waseda.jp)

Tatsuya Nakata
Rikkyo University, Japan (nakata@rikkyo.ac.jp)

Xuehong (Stella) He
Swansea University, UK (xuehong.he@swansea.ac.uk)

**Abstract**

The aim of this study was to contribute to the field of computer-assisted language learning (CALL) by investigating the individualization of intentional vocabulary learning. A total of 118 Japanese-speaking university students studied 20 low-frequency English words using flashcard software over two learning sessions. The participants practiced retrieval of vocabulary under different learning schedules, with short or long time intervals between encounters of the same word in each learning session: Short–Short, Short–Long, Long–Short, and Long–Long. Two individual difference measures – learning efficiency and language aptitude – were examined as predictors of long-term second language (L2) vocabulary retention. Learning efficiency was operationalized as the number of trials needed to reach a learning criterion in each session, whereas a component of aptitude (rote memory ability) was measured by a subtest of Language Aptitude Battery for the Japanese. Multiple regression and dominance analyses were conducted to evaluate the relative importance of learning efficiency and language aptitude in predicting delayed vocabulary posttest scores. The results revealed that learning efficiency in the second learning session was the strongest predictor of vocabulary retention. Language aptitude, however, did not significantly predict vocabulary retention. Moreover, the predictive power of learning efficiency increased when the data were analyzed within each learning schedule, underscoring the need to assess learners' abilities under specific learning conditions for optimizing their computer-assisted learning performance. These findings not only inform the development of more effective, individualized CALL systems for L2 acquisition but also emphasize the importance of gauging individuals' abilities such as learning efficiency in a more flexible, context-sensitive manner.

**Keywords:** vocabulary learning; individualized CALL; language aptitude; learning efficiency; practice distribution

## 1. Introduction

Acquiring extensive vocabulary knowledge is a formidable task for second language (L2) learners aiming to achieve advanced proficiency. The challenge often necessitates mastering thousands of vocabulary items, including single words as well as multi-word items that are essential for proficient reading, listening, writing, and speaking skills (e.g. Nation, 2006; Wray, 2002). Extant research in this context indicates that incidental and intentional vocabulary learning strategies are

**Cite this article:** Suzuki, Y., Nakata, T. & He, X.(Stella). (2025). Efficiency trumps aptitude: Individualizing computer-assisted second language vocabulary learning. *ReCALL* 37(1): 96–113. https://doi.org/10.1017/S095834402400020X

equally important in building a robust lexical knowledge base. While incidental learning serves as a major pathway for acquiring various aspects of lexical knowledge, such as meaning, form, and use (Nation, 2022), intentional vocabulary practice, such as the use of flashcards (also referred to as word cards), is one of the most efficient methods for learning the form–meaning connections among lexical items (for recent reviews, see Nakata, 2020; Webb, Yanagisawa & Uchihara, 2020).

Advances in technology have introduced computer-assisted language learning (CALL) and mobile-assisted language learning (MALL) systems into the L2 domain, which have shown the potential to accelerate intentional vocabulary acquisition (e.g. Lee, Warschauer & Lee, 2019; Li & Hafner, 2022; Nakata, 2011; Seibert Hanson & Brown, 2020; Stockwell, 2022; Yu & Trainin, 2022).[1] A key feature of these computer-assisted vocabulary learning systems is a learning criterion that determines when to remove vocabulary items from the practice list. Kaitsu (2023), for instance, surveyed six commercially available flashcard apps for smartphones and found that in four apps, the criterion for excluding target vocabulary items from further practice (e.g. after one correct response) is predetermined without considering individual learner performance. However, the optimal number of trials required for long-term retention can vary significantly among learners (Zerr *et al.*, 2018). This individual variability underscores the potential for tailoring the number of practice trials in computer-assisted vocabulary learning based on individual learner needs and abilities.

Expanding this idea of CALL system customization, the learning abilities of individuals can serve as a rich data source for adapting the amount and types of practice within CALL activities (Pawlak & Kruk, 2022; Ruiz, Rebuschat & Meurers, 2023). For instance, to achieve the same level of performance, more capable learners (e.g. those with greater memory capacity) may require fewer learning trials through CALL than those with lower abilities.

Two primary approaches have been proposed to operationalize these individual abilities in the context of computer-assisted vocabulary learning. The first focuses on *language aptitude*, which influences the rate of vocabulary acquisition as well as its retention. Accordingly, in the aptitude-treatment interaction (ATI) framework, which aims to match learning conditions with language aptitude, CALL is recognized as a particularly promising field for conducting aptitude research and applying the obtained findings (e.g. DeKeyser, 2019; Malone, 2018; Ruiz, Rebuschat & Meurers, 2021). The second approach involves real-time collection of learning data during CALL activities to derive a *learning-efficiency score*, typically reflecting the time learners require to master target knowledge. As this information can be gathered during CALL activities, a learning-efficiency score can be a useful index for predicting the retention of learned information, including L2 vocabulary knowledge (Zerr *et al.*, 2018).

The current study aims to contribute to the extant CALL research by focusing on the personalization of intentional vocabulary learning. Specifically, its goal is to examine both language aptitude and learning-efficiency score as the potential predictors of vocabulary acquisition and retention. The obtained findings will inform the design of more effective adaptive computer-assisted vocabulary learning systems tailored to the needs and abilities of individual learners.

## 2. Literature review

### 2.1 Language aptitude and vocabulary learning

Language aptitude has long been a focus of L2 acquisition research (see Chalmers, Eisenchlas, Munro & Schalley, 2021, for a recent overview). Research syntheses such as those provided by Li (2015, 2016) underscore considerable individual variations in learning rates and ultimate attainment of various L2 skills. While the role of aptitude in the acquisition of L2 grammar and higher-order skills (reading, listening, writing, and speaking) has been extensively studied, research focusing on

[1] For brevity, we use CALL to cover both CALL and MALL systems for L2 learning.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

vocabulary learning remains comparatively marginal. This gap in the literature is surprising, given that individual differences also influence the vocabulary learning rate and retention.

Indeed, findings yielded by a few empirical studies in this domain indicate that language aptitude does play a role in both intentional and incidental vocabulary acquisition. Specifically, language aptitude predicts intentional vocabulary learning through deliberate (sub)vocal rehearsal (Dahlen & Caldwell-Harris, 2013) as well as facilitates incidental vocabulary learning through activities such as reading (Malone, 2018) and video viewing (Muñoz, Pattemore & Avello, 2022; Suárez & Gesa, 2019).

As language aptitude is multi-componential, it has been traditionally conceptualized as consisting of phonetic coding ability, rote memory ability, inductive learning ability, and grammatical sensitivity (Carroll & Sapon, 1959). Among these constructs, rote memory ability appears to be the most pertinent to intentional vocabulary learning. Therefore, assessing an individual's aptitude such as rote memory ability through prior testing could serve as a basis for customizing the vocabulary learning parameters in CALL.

Computer-assisted vocabulary learning can be personalized by manipulating two major parameters: the amount and/or distribution of practice. First, the amount of vocabulary practice required may differ depending on learners' abilities. In recent L2 grammar research, Serfaty and Serrano (2024) aimed to identify the optimal amount of L2 grammar practice using computerized flashcards (translation from first language [L1] to target language). For this purpose, the authors subjected the study participants to different amounts of grammar practice on an online learning platform and assessed the retention of their grammar knowledge via a two-week delayed posttest. The findings suggest that some learners needed a significantly greater number of practice trials than others to achieve durable grammar knowledge. Although Serfaty and Serrano (2024) did not measure the language aptitude of their participants, it is plausible to assume that their different learning rates and different levels of grammar knowledge retention were influenced by variations in their individual aptitudes.

The grammar practice results obtained by Serfaty and Serrano (2024) should be relevant to the context of vocabulary retrieval practice, where individuals recall information about target lexical items from memory. Retrieval practice is known to enhance long-term retention (Karpicke & Roediger, 2008; Nakata, 2017; Seibert Hanson & Brown, 2020), and its optimal amount may depend on the learners' aptitude levels. High-aptitude learners may require fewer retrieval opportunities to achieve optimal learning outcomes compared to their low-aptitude counterparts. Consequently, understanding the interplay between aptitude and vocabulary learning can offer significant insights for personalizing vocabulary learning in CALL.

Second, the distribution of practice can also be manipulated to personalize computer-assisted vocabulary learning, given that temporal spacing between practice opportunities for a given item influences long-term retention (see Kim & Webb, 2022, for a recent meta-analysis). For instance, a longer-spacing schedule (e.g. practicing a given word at 20-minute intervals) typically leads to better long-term retention than a massed schedule (practicing a given word multiple times in a row) or shorter-spacing schedule (e.g. practicing a given word every 30 seconds; e.g. Nakata & Webb, 2016). Yet the moderating role of language aptitude in the effectiveness of different vocabulary learning schedules remains insufficiently studied. Findings yielded by L2 grammar learning research, however, suggest that language aptitude may be a significant predictor of retention, especially when the treatment involves long temporal spacing between repetitions (e.g. Suzuki, 2019; Suzuki & DeKeyser, 2017). This effect is posited to exist because learners with higher aptitude can retain the previously learned information (i.e. target grammatical rules) until the next session. Conversely, for low-aptitude learners, memory of previously studied grammatical structures may decay completely during this time, which may suggest that these learners would benefit from shorter spacing. As ATI research has predominantly focused on grammar learning, and meaningful ATI patterns are rarely reported in the existing vocabulary research, the role of

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

language aptitude in determining the optimal amount and distribution of vocabulary practice requires further investigation.

### 2.2 Learning efficiency and vocabulary learning

Despite the burgeoning L2 vocabulary research, Nation (2013) pointed out that the link between learning efficiency during the training phase and subsequent knowledge retention remains largely unexplored. There are divergent predictions regarding this relationship. On the one hand, learners requiring a greater number of trials could potentially experience better long-term retention. Such an outcome could be explained by the desirable difficulty framework (Bjork, 1994; Suzuki, Nakata & DeKeyser, 2019), positing that the additional cognitive burden imposed by a greater number of trials during the learning phase (which are indicative of slow, effortful learning) can facilitate deeper encoding, thereby improving retention. On the other hand, learners who require fewer trials may be more efficient and thus may possess higher language aptitude, including rote memory ability, typically measured via tasks that require quick and accurate recall of vocabulary lists. Accordingly, it is conceivable that these learners may exhibit better long-term retention.

The latter viewpoint is supported by the study conducted by Zerr *et al.* (2018) in the field of cognitive psychology. As a part of their study, English-speaking participants studied 45 Lithuanian–English word pairs. After seeing all target Lithuanian words accompanied by their English translations at the outset of the training phase, the participants engaged in multiple retrieval practice rounds until they correctly recalled each of the target items. In other words, correctly recalled pairs were dropped from subsequent retrieval practice trials. After a 5-minute delay introduced by a distractor task, the participants took an immediate receptive posttest, which required them to type the meaning of the 45 Lithuanian words in English. A learning-efficiency score was computed for each participant by combining the accuracy score for the initial practice round, the number of trials required to reach the learning criterion, and the accuracy score achieved on the immediate posttest. The results showed that those with higher efficiency scores tended to perform better on the immediate posttest.

Although the study conducted by Zerr *et al.* (2018) was not originally designed to inform CALL optimization, the obtained findings are valuable because they suggest that learning-efficiency scores may be used to personalize and optimize computer-assisted vocabulary learning. For instance, for learners with high efficiency scores, flashcard software could terminate the treatment after all target items are recalled correctly once. Those with low efficiency scores, in contrast, may be required to practice until a stricter criterion (e.g. two or more correct answers) has been attained.

Nonetheless, the research design adopted by Zerr *et al.* (2018) could be further improved to assist in the personalization of computer-assisted vocabulary learning (Pawlak & Kruk, 2022). First, although multiple learning sessions are usually required to increase the likelihood of vocabulary retention, only a single learning session was included in the treatment protocol. Second, the lack of a delayed posttest makes it difficult to evaluate the role of learning efficiency in long-term retention. Third, the calculation of the learning-efficiency score included an immediate posttest score, which could have inflated their correlation between the learning-efficiency and immediate posttest scores.[2] In the current study, these limitations were addressed by (a) requiring participants to learn vocabulary in two sessions distributed over weeks, (b) administering a delayed posttest seven days after the second learning session, and (c) operationalizing learning efficiency as the number of trials required to complete the first and second training phases.

[2] Zerr *et al.* (2018) stated that the immediate posttest score was included in the calculation of overall measure of learning efficiency because it was correlated with the accuracy score learners attained in the initial practice round and the number of trials they needed to reach the specified criterion.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

Learning efficiency is presumably related to language aptitude, given a documented association between individual cognitive abilities, such as IQ and processing speed, and the learning-efficiency score (Zerr *et al.*, 2018). However, several notable differences seem to exist between learning-efficiency and aptitude scores. First, an aptitude score must be measured independently and prior to CALL activities. In contrast, the learning-efficiency score is measured during CALL activities and can even be updated in situ as learners engage in vocabulary learning. For instance, when vocabulary learning spans across multiple days, the learning-efficiency score derived from the first session can be used to adjust the parameters for the second session, and the score from the second session can be employed for subsequent sessions. In other words, while aptitude is considered a static or stable personal trait (Carroll, 1981; Doughty, 2019), the learning-efficiency score is a more dynamic state that can be adjusted during the ongoing learning process. The reconfiguration of learners' ability or aptitude situated within and leveraged by CALL systems may thus resonate with a more flexible and dynamic view of individual differences in L2 acquisition theories such as complex dynamic systems (Larsen-Freeman & Cameron, 2008; Lowie & Verspoor, 2019). Although no research has been conducted to compare the predictive power of aptitude and learning-efficiency scores, such investigations could be invaluable for informing optimal vocabulary practice in CALL systems, as well as the broader discourse on ATI in L2 research.

## 3. Current study

The goal of the present study was to evaluate the usefulness of learning-efficiency and aptitude scores as predictors of vocabulary learning in CALL. The data from a study previously published by Nakata, Suzuki and He (2023) that focused on the role of practice distribution were reanalyzed. In this prior study, Japanese-speaking university students learned English–Japanese word pairs in CALL classrooms with different time lags over two learning sessions (i.e. short- versus long-spaced schedules within each learning session). They subsequently took part in a seven-day delayed posttest. The goal of the original study was to investigate the effects of different time lags over two learning sessions on long-term vocabulary retention. The reanalysis reported here focuses on individual difference factors in vocabulary learning that were not examined in the original study. Specifically, the aim was to examine the relative importance of learning efficiency (i.e. the number of retrieval trials required to complete the training phase) and aptitude (i.e. rote learning ability) for long-term vocabulary retention. The following two research questions (RQs) were addressed:

1. To what extent do learning efficiency and language aptitude predict the retention of L2 vocabulary?
2. Do different vocabulary learning schedules moderate the predictive power of learning efficiency and language aptitude?

## 4. Method

As noted in the previous section, as a part of this study, the data obtained by Nakata *et al.* (2023) were reanalyzed. For clarity, a brief overview of the methodology adopted in the original study is given below.

### *4.1 Participants*

The initial pool of participants consisted of 170 Japanese-speaking university students. Because the purpose of the original study was to examine the effects of spacing schedule, the participants were assigned to one of the following groups: Short–Short, Short–Long, Long–Short, and Long–Long, reflecting the spacing schedule adopted for the first and second treatment session, respectively (see the Procedure section for details).

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

While 170 students were recruited at the outset, the data related to 52 individuals were excluded from the final analysis because (a) they did not have scores on the TOEIC test, (b) they missed at least one of the three experimental sessions, (c) they stated in the post-study survey that they were not L1 speakers of Japanese, or (d) they stated in the post-study survey that they had been exposed to at least one of the target items outside the study. Among the remaining 118 students, 33, 30, 30, and 25 belonged to the Short–Short, Short–Long, Long–Short, and Long–Long groups, respectively. Based on their TOEIC scores ($M = 489.3$, $SD = 144.7$), the participants' English proficiency levels fell between the A2 (elementary) and B1 (intermediate) levels in the Common European Framework of Reference for Languages (CEFR) benchmark. No significant difference in the TOEIC test scores was detected among the four groups, $F(3, 156) = 0.03$, $p = .99$, $\eta^2_p < 0.01$.

### 4.2 Materials

The following 20 low-frequency English words were used as target items: *apparition*, *billow*, *cadge*, *citadel*, *dally*, *fawn*, *fracas*, *gouge*, *grig*, *levee*, *loach*, *mane*, *mirth*, *nadir*, *pique*, *quail*, *rue*, *scowl*, *toupee*, and *warble*. All chosen words fall at the 10,000 word level or below, according to the British National Corpus frequency lists. These particular words were selected based on prior research indicating their likely unfamiliarity to most Japanese undergraduate students (Nakata & Webb, 2016). The set comprised 20 items, with a distribution of 12 nouns and eight verbs. This distribution mirrors the approximate 6:4 noun-to-verb ratio found in natural text (Webb, 2005).

### 4.3 Procedure

The study was conducted during three weekly English language classes at a Japanese university. The experiment was administered on computers using Gorilla Experiment Builder (https://app.gorilla.sc/).

#### 4.3.1 Session 1

At the outset of the first experimental session, the students read information sheets about the study and electronically signed the informed consent form. Before the vocabulary learning session, the paired-associate section of Language Aptitude Battery for the Japanese (LABJ; Sasaki, 1993) was administered. This section of LABJ is a Japanese-translated version of Part V of the Modern Language Aptitude Test and is intended to assess test-takers' ability to learn L1 and L2 word pairs in a decontextualized format. When taking this test, students were presented with a list of 24 Kurdish–Japanese word pairs and were instructed to remember as many items as possible in 2 minutes. Next, the participants took a 4-minute multiple-choice receptive test, in which they were required to choose the correct translation of a target Kurdish word from five Japanese options.

After completing the LABJ, the students practiced using the flashcard program with four sample items. The practice phase was followed by productive and receptive pretests. The former required the students to translate Japanese (L1) words into English (L2), whereas the latter required them to translate English target words into Japanese.

The first treatment session commenced immediately after the pretests and involved up to five encounters of the 20 target English words. In the first encounter, the target English word appeared on the screen for 8 seconds together with its Japanese translation (presentation trials). From the second encounter, target items were practiced in a form recall format (retrieval trials). Specifically, students were presented with the Japanese translation of a target item and were instructed to type the corresponding English word. Each response was followed by feedback, where the correct target word, together with its Japanese translation, appeared on the screen for 5 seconds. Items that elicited correct responses were excluded from the list used in further practice.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

In the first treatment session, the participants in the Short–Short and Short–Long groups followed a short-spaced schedule, whereas a long-spaced schedule was adopted for those in the Long–Short and Long–Long groups. In the short-spaced schedule, encounters of a given target item were separated by three other items on average, whereas in the long-spaced schedule, encounters for a given target item were separated by 19 other items on average (unless other target items had been excluded from practice after eliciting a correct response). The first treatment session ended when (a) all target items were answered correctly once, or (b) the target items not answered correctly underwent four retrieval trials.

#### 4.3.2 Session 2

The second treatment session was conducted one week after the first session and was based on the identical protocol, except that all trials were retrieval trials, and there were no presentation trials. In this session, the participants in the Short–Short and Long–Short groups studied with a short-spaced schedule, and those in the Short–Long and Long–Long groups studied with a long-spaced schedule.

#### 4.3.3 Session 3

One week after the second experimental session, the students participated in the third session, which consisted of productive and receptive posttests (administered in that order), and a questionnaire. The first posttest required the students to translate Japanese (L1) words into English (L2), whereas the second required them to translate English target words into Japanese. Upon completion of the posttests, the participants were asked to complete a questionnaire probing into their perceptions of the experiment.

### 4.4 Statistical analysis

In order to assess the relative importance of predictors (the learning-efficiency scores in Session 1 and 2 [LE1 and LE2, respectively] and LABJ score), multiple regression analyses were conducted with the delayed productive and receptive posttest score as a dependent variable, respectively, in the entire sample (RQ1) and for four different spacing groups (RQ2). Multicollinearity, as the key statistical assumption that must be addressed for multiple regression, was assessed using the variance inflation factor values, which were below 10, thereby indicating the absence of multicollinearity issues. To evaluate the relative importance of predictors by minimizing their suppression effect (see Mizumoto, 2023, for details), dominance analysis was further conducted using an R-based web application (https://langtest.jp/shiny/relimp/). Dominance weights were calculated to compare the relative effect sizes of predictors.

In order to further probe the interaction between group and predictors (RQ2), generalized linear mixed-effects models were constructed for the receptive and productive posttest scores, respectively. Predictors were LE1, LE2, and LABJ. The interaction terms with group were also created with each predictor. Details of the full analyses are reported in Appendix A in the supplementary material. The final model is presented below:

$$\begin{aligned}\text{Posttest Score} \sim\ & \text{Group} + \text{LE1} + \text{LE2} + \text{LABJ} + \text{Group} : \text{LE1} + \text{Group} : \text{LE2} + \text{Group} : \text{LABJ} \\ & + (1|\text{ID}) + (1 + \text{Group} + \text{LE1} + \text{LE2} + \text{LABJ} + \text{Group} : \text{LE1} + \text{Group} : \text{LE2} \\ & + \text{Group} : \text{LABJ}|\text{Item})\end{aligned}$$

A significant interaction between group and predictor ($p < .05$) indicates that the importance of predictor differs depending on group or learning schedules.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Table 1.** Descriptive statistics for the delayed posttests

| | Productive test | | | Receptive test | | |
|---|---|---|---|---|---|---|
| | *M* | 95% CI | *SD* | *M* | 95% CI | *SD* |
| Short–Short | 19.6% | [13.7%, 25.4%] | 17.2% | 44.9% | [37.3%, 52.4%] | 22.1% |
| Short–Long | 21.7% | [15.9%, 27.4%] | 16.0% | 52.2% | [45.1%, 59.2%] | 19.8% |
| Long–Short | 36.7% | [28.8%, 44.5%] | 21.9% | 65.2% | [59.0%, 71.4%] | 17.3% |
| Long–Long | 48.2% | [42.0%, 55.4%] | 18.5% | 70.4% | [62.3%, 78.5%] | 20.6% |
| Total | 30.5% | [26.6%, 34.4%] | 21.5% | 57.3% | [53.2%, 61.3%] | 22.2% |

*Note*. CI = confidence interval.

## 5. Results

### 5.1 Posttest scores

The Cronbach alpha coefficient of .83 was obtained for both productive and receptive posttests, which was within the acceptable range, as previously established in a similar outcome task (e.g. Nakata & Webb, 2016). As can be seen from the descriptive statistics of the delayed posttest scores presented in Table 1, some variances (indicated by the *SD* values, which averaged at about 20%) suggest considerable differences in the performance among learners. Yet, as reported by Nakata *et al.* (2023), inferential statistics (ANCOVA) revealed a clear order of effectiveness, namely Long–Long > Long–Short > Short–Long = Short–Short for the productive posttest and Long–Long = Long–Short > Short–Long = Short–Short for the receptive posttest.

### 5.2 Aptitude and learning-efficiency scores

The Cronbach alpha coefficient for the LABJ was .89, indicating a high internal consistency. The descriptive statistics for the learning-efficiency and LABJ scores are presented in Table 2. As the learning-efficiency score was defined as the number of retrieval trials required for participants to reach the criterion of one correct response, lower values reflect higher efficiency.[3]

One-way ANOVAs were conducted to examine the potential group difference(s) with respect to each score. The main effect of group assignment was significant on the learning-efficiency score in both sessions, Session 1: $F(3, 114) = 30.08$, $p < .001$, $\eta^2_p = 0.44$; Session 2: $F(3, 114) = 22.23$, $p < .001$, $\eta^2_p = 0.37$. When post hoc comparisons with Holm correction were conducted for the learning-efficiency score in Session 1, the findings indicated that two groups that started with the long-spaced schedule (i.e. Long–Long and Long–Short) required a significantly larger number of trials than the Short–Short and Short–Long groups. Based on the learning-efficiency score in Session 2, the number of trials needed to achieve the learning criterion (one correct response) differed across the four groups as follows (from largest to smallest): Short–Long > Long–Long = Short–Short > Long–Short. On the other hand, there was no significant group difference on the LABJ score, $F(3, 114) = 0.37$, $p = .78$, $\eta^2_p = 0.01$.

Correlations between the learning-efficiency and LABJ scores were of medium magnitude in both sessions, Session 1: $r = -.31$, 95% CI [−0.48, −0.15], $p < .001$; Session 2: $r = -.43$, 95% CI

[3] It may appear surprising that participants required a larger number of trials to reach the same criterion in Session 2 than in Session 1. This may be explained by the slightly different procedures in the two sessions. Specifically, in Session 1, each target item was introduced during presentation trials, where learners were presented with L1–L2 word pairs. In contrast, all trials were retrieval trials in Session 2 (see Procedure). Accordingly, the opportunity to gain familiarity with the word pairs at the outset of Session 1 likely contributed to a greater number of correct responses during initial retrieval attempts, due to which fewer trials were required to complete the first treatment.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Table 2.** Descriptive statistics for learning-efficiency and LABJ scores

| Learning-efficiency score (Session 1) | *M* | 95% CI | *SD* |
|---|---|---|---|
| Short–Short | 36.76 | [33.85, 40.09] | 9.23 |
| Short–Long | 34.53 | [31.67, 37.43] | 8.23 |
| Long–Short | 54.20 | [50.26, 58.20] | 11.58 |
| Long–Long | 52.12 | [48.00, 56.12] | 11.24 |
| Total | 43.88 | [41.29, 46.29] | 13.33 |
| **Learning-efficiency score (Session 2)** | *M* | 95% CI | *SD* |
| Short–Short | 50.30 | [48.06, 52.64] | 7.10 |
| Short–Long | 60.43 | [58.23, 62.83] | 6.56 |
| Long–Short | 44.83 | [41.90, 47.63] | 8.39 |
| Long–Long | 50.40 | [47.20, 53.44] | 8.13 |
| Total | 51.51 | [49.76, 53.02] | 9.38 |
| **LABJ** | *M* | 95% CI | *SD* |
| Short–Short | 21.27 | [19.97, 22.40] | 3.61 |
| Short–Long | 20.20 | [18.47, 21.77] | 4.75 |
| Long–Short | 20.60 | [19.03, 22.17] | 4.46 |
| Long–Long | 20.32 | [18.28, 22.12] | 4.98 |
| Total | 20.63 | [19.71, 21.45] | 4.40 |

*Note*. The maximum LABJ score was 24. CI = confidence interval; LABJ = Language Aptitude Battery for the Japanese (Sasaki, 1993).

[−0.60, −0.24], $p < .001$. This finding suggests that learning-efficiency scores were related to but distinct from rote memory ability measured by the LABJ subtest. Still, it is worth noting that, while the strength of the relationship between the learning-efficiency and LABJ score did not vary greatly across groups, slightly greater magnitude was obtained for Session 2 ($-.59 \leq r \leq -.40$) than for Session 1 ($-.47 \leq r \leq -.36$).

### 5.3 Multiple regression and relative dominance analyses

A multiple regression analysis was conducted to investigate the relative importance of the learning-efficiency and LABJ scores for the delayed posttest scores (RQ1). The omnibus models were significant in both productive and receptive posttests, productive: $F(3, 114) = 27.51$, $p < .001$, $R^2 = 0.42$; receptive: $F(3, 114) = 23.41$, $p < .001$, $R^2 = 0.38$. As shown in Table 3, the learning-efficiency score in Session 2 was the only significant predictor in both models. The dominance analysis indicated that the learning-efficiency score in Session 2 accounted for 39% of the variance in the productive posttest performance and 33% of the variance in the receptive posttest outcome. The LABJ score, in contrast, only accounted for 2% and 5% of the variance, respectively.

In order to investigate the extent to which the predictive power of learning-efficiency and LABJ scores is moderated by vocabulary learning schedules, multiple regression analyses were conducted for each group (RQ2) and the results are presented in Tables 4 and 5. All models were statistically significant, accounting for larger variance than the full-sample models ($R^2 \geq .52$), except for the model related to the productive posttest performance in the Short–Long group

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Table 3.** Relative importance of predictors for the delayed posttest performance

| | Relative importance (%) | 95% CI | Estimate | *SE* | *t* | *p* |
|---|---|---|---|---|---|---|
| Productive posttest | | | | | | |
| (Intercept) | | | 23.17 | 3.42 | 6.77 | .00 |
| Learning efficiency Session 1 | 0.00 (0%) | [0.00, 0.00] | 0.01 | 0.02 | 0.35 | .73 |
| Learning efficiency Session 2 | 0.39 (94%) | [0.25, 0.52] | −0.31 | 0.04 | −8.54 | .00 |
| LABJ | 0.02 (6%) | [0.01, 0.06] | −0.07 | 0.08 | −0.86 | .39 |
| Receptive posttest | | | | | | |
| (Intercept) | | | 24.82 | 3.65 | 6.81 | .00 |
| Learning efficiency Session 1 | 0.00 (1%) | [0.00, 0.01] | 0.00 | 0.03 | −0.19 | .85 |
| Learning efficiency Session 2 | 0.33 (87%) | [0.20, 0.45] | −0.28 | 0.04 | −7.21 | .00 |
| LABJ | 0.05 (13%) | [0.01, 0.13] | 0.06 | 0.09 | 0.68 | .50 |

*Note*. Correlation coefficients are presented in Appendix B in the supplementary material. CI = confidence interval; LABJ = Language Aptitude Battery for the Japanese (Sasaki, 1993).

**Table 4.** Results of multiple regression analyses by groups and posttests

| Groups | Posttests | *df1, df2* | *F* | *p* | $R^2$ |
|---|---|---|---|---|---|
| Short–Short | Productive | 3, 29 | 12.72 | < .001 | .57 |
| | Receptive | 3, 29 | 10.57 | < .001 | .52 |
| Short–Long | Productive | 3, 26 | 3.51 | .03 | .29 |
| | Receptive | 3, 26 | 9.60 | < .001 | .53 |
| Long–Short | Productive | 3, 26 | 24.09 | < .001 | .74 |
| | Receptive | 3, 26 | 12.43 | < .001 | .59 |
| Long–Long | Productive | 3, 21 | 12.98 | < .001 | .65 |
| | Receptive | 3, 21 | 12.36 | < .001 | .64 |

($R^2 = .29$). Overall, the patterns observed in the results were consistent with those yielded by the full-sample analyses. The learning-efficiency score in Session 2 was the only significant predictor in all models. Once again, the model pertaining to the productive posttest performance in the Short–Long group was the only exception (but even in this case, the results were close to the statistical significance, $p = .055$).

The amount of variance explained by the learning-efficiency score in the productive test performance as a part of Session 2 was nearly 50% for the Short–Short, Long–Short, and Long–Long groups. Therefore, with the exception of the Long–Short group, the learning-efficiency score exhibited a higher predictive power with respect to knowledge retention within each group than in the full sample. Another notable difference from the full-sample analysis is that the learning-efficiency score in Session 1 accounted for the variance of 9%–20%, whereas in the full-sample model, its contribution was negligible (0%).

The LABJ score was not a significant predictor in any of the eight regression models developed as a part of this investigation, and the variance it explained was below 10% in all cases except the model related to the Long–Long group. For the receptive test score in this group, the LABJ score accounted for 23% of variance, exceeding 11% explained by the learning-efficiency score in

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Table 5.** Relative importance of predictors for the delayed posttest performance by groups

| Short–Short | Relative importance (%) | 95% CI | Estimate | *SE* | *t* | *p* |
|---|---|---|---|---|---|---|
| Productive posttest | | | | | | |
| (Intercept) | | | 22.16 | 5.03 | 4.40 | .00 |
| Learning efficiency Session 1 | 0.09 (16%) | [0.00, 0.20] | 0.08 | 0.07 | 1.24 | .23 |
| Learning efficiency Session 2 | 0.45 (79%) | [0.28, 0.66] | −0.43 | 0.08 | −5.09 | .00 |
| LABJ | 0.03 (5%) | [−0.03, 0.14] | 0.02 | 0.13 | 0.13 | .90 |
| Receptive posttest | | | | | | |
| (Intercept) | | | 27.96 | 6.80 | 4.11 | .00 |
| Learning efficiency Session 1 | 0.17 (33%) | [0.07, 0.41] | −0.10 | 0.09 | −1.05 | .30 |
| Learning efficiency Session 2 | 0.30 (57%) | [0.08, 0.56] | −0.34 | 0.11 | −2.94 | .01 |
| LABJ | 0.05 (10%) | [0.01, 0.21] | 0.07 | 0.18 | 0.38 | .71 |
| Short–Long | Relative importance (%) | 95% CI | Estimate | *SE* | *t* | *p* |
| Productive posttest | | | | | | |
| (Intercept) | | | 22.71 | 7.62 | 2.98 | .01 |
| Learning efficiency Session 1 | 0.11 (39%) | [0.02, 0.23] | −0.10 | 0.08 | −1.37 | .18 |
| Learning efficiency Session 2 | 0.16 (56%) | [0.03, 0.41] | −0.21 | 0.10 | −2.01 | .05 |
| LABJ | 0.02 (5%) | [0.00, 0.03] | −0.11 | 0.13 | −0.80 | .43 |
| Receptive posttest | | | | | | |
| (Intercept) | | | 44.55 | 7.66 | 5.81 | .00 |
| Learning efficiency Session 1 | 0.17 (32%) | [0.03, 0.37] | −0.14 | 0.08 | −1.90 | .07 |
| Learning efficiency Session 2 | 0.33 (62%) | [0.11, 0.56] | −0.40 | 0.10 | −3.77 | .00 |
| LABJ | 0.03 (7%) | [0.00, 0.07] | −0.26 | 0.13 | −1.94 | .06 |
| Long–Short | Relative importance (%) | 95% CI | Estimate | *SE* | *t* | *p* |
| Productive posttest | | | | | | |
| (Intercept) | | | 35.56 | 4.59 | 7.75 | .00 |
| Learning efficiency Session 1 | 0.20 (27%) | [0.07, 0.35] | −0.08 | 0.05 | −1.70 | .10 |
| Learning efficiency Session 2 | 0.50 (68%) | [0.36, 0.67] | −0.42 | 0.07 | −6.01 | .00 |
| LABJ | 0.04 (5%) | [0.00, 0.11] | −0.23 | 0.11 | −2.05 | .05 |
| Receptive posttest | | | | | | |
| (Intercept) | | | 25.89 | 4.51 | 5.74 | .00 |
| Learning efficiency Session 1 | 0.09 (15%) | [0.01, 0.25] | 0.00 | 0.05 | 0.07 | .95 |
| Learning efficiency Session 2 | 0.43 (72%) | [0.29, 0.59] | −0.31 | 0.07 | −4.45 | .00 |
| LABJ | 0.08 (13%) | [0.01, 0.23] | 0.04 | 0.11 | 0.34 | .74 |
| Long–Long | Relative importance (%) | 95% CI | Estimate | *SE* | *t* | *p* |
| Productive posttest | | | | | | |
| (Intercept) | | | 28.69 | 5.49 | 5.22 | .00 |
| Learning efficiency Session 1 | 0.10 (15%) | [0.02, 0.20] | 0.06 | 0.06 | 1.04 | .31 |

(*Continued*)

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Table 5.** (*Continued*)

| | | | | | | |
|---|---|---|---|---|---|---|
| Learning efficiency Session 2 | 0.47 (72%) | [0.23, 0.65] | −0.43 | 0.09 | −4.72 | .00 |
| LABJ | 0.09 (13%) | [0.01, 0.27] | −0.03 | 0.12 | −0.23 | .82 |
| Receptive posttest | | | | | | |
| (Intercept) | | | 22.61 | 6.22 | 3.63 | .00 |
| Learning efficiency Session 1 | 0.11 (17%) | [0.02, 0.21] | 0.00 | 0.07 | 0.01 | 1.00 |
| Learning efficiency Session 2 | 0.30 (47%) | [0.11, 0.44] | −0.28 | 0.10 | −2.72 | .01 |
| LABJ | 0.23 (36%) | [0.06, 0.41] | 0.28 | 0.13 | 2.06 | .05 |

*Note*. Correlation coefficients, along with scatterplots, are presented in Appendix B in the supplementary material. CI = confidence interval; LABJ = Language Aptitude Battery for the Japanese (Sasaki, 1993).

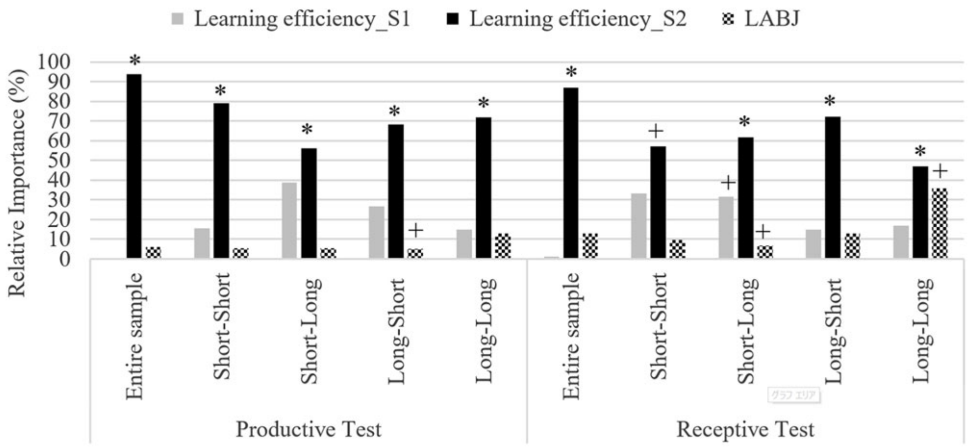

**Figure 1.** Relative importance of predictors in multiple regression models.
*Note.* An asterisk (*) indicates a significant predictor ($p < .05$), whereas a cross (+) denotes a marginally significant predictor in the model ($p < .10$).

Session 1, while being below 30% explained by the learning-efficiency score in Session 2. Figure 1 summarizes the relative importance of learning-efficiency and aptitude scores in predicting vocabulary retention across 10 regression models.

Finally, generalized linear mixed-effects modeling was conducted for productive and receptive posttest scores. For the productive posttest, three significant interactions were detected (for full results, see Appendix C in the supplementary material). The interactions between group and LE1 were significant in the following comparisons: Short–Short versus Short–Long ($z = 2.40$, $p = .02$) and Long–Long versus Short–Long ($z = 2.05$, $p = .04$). These interactions suggest that the importance of the learning-efficiency score at Session 1 was greater in the Short–Long group than in the Long–Long or Short–Short group. The interaction between group and LE2 was significant for the comparison between Short–Short and Short–Long ($z = -2.42$, $p = .02$). This suggests that the learning-efficiency score at Session 2 was less important in the Short–Long group than in the Short–Short group, which in turn corroborates the larger role of the learning-efficiency score at Session 1 in the Short–Long group. For the receptive posttest, only one significant interaction between group and LABJ was revealed (for full results, see Appendix D in the supplementary material). This suggests that the importance of the LABJ score was greater in the Long–Long group than in the Short–Long group ($z = 2.97$, $p = .003$).

Taken together, the findings from generalized linear mixed-effects modeling highlighted more nuanced and complex patterns beyond the overall importance of the learning-efficiency score at Session 2. Specifically, for the productive posttest, the importance of the learning-efficiency score at Session 1 was prominent in the Short–Long group; for the receptive posttest, the LABJ score was prominent in the Long–Long group (see Figure 1).

## 6. Discussion

The first RQ addressed in this work pertained to the capacity of learning-efficiency and aptitude scores to predict vocabulary acquisition irrespective of the spacing schedule. In the corresponding models, the learning-efficiency score in Session 2 emerged as the only significant predictor of vocabulary knowledge retention. According to the dominance analysis results, the learning-efficiency score in Session 2 accounted for 39% of the variance in the productive posttest performance (constituting 94% of the total variance explained), and 33% of the variance in the receptive posttest performance (constituting 87% of the total variance explained). Conversely, by accounting for only 2% (6% of the total variance explained) and 5% (13% of the total variance explained) of the productive and receptive test performance, respectively, aptitude scores were not significant predictors of vocabulary knowledge retention. As the learning-efficiency score in Session 1 did not account for any variance in either posttest, these findings suggest that the learning-efficiency score in Session 2 was the most informative variable for predicting the retention of L2 vocabulary. This observation also corroborates the recent findings obtained by Zerr *et al.* (2018) in their cognitive psychology study, indicating that faster learners – despite spending less time on practice – can acquire more durable knowledge. Accordingly, the learning-efficiency score, such as the one derived from the number of trials during CALL, is an easily measurable index that can be a valuable indicator of other types of L2 learning (e.g. pronunciation, grammar, pragmatics) through CALL systems, which can be investigated in future CALL-based research.

The mechanisms underpinning the faster learners' capacity to retain vocabulary knowledge for a longer period could involve their superior language aptitude. This argument is supported by the significant positive correlation between learning efficiency and language aptitude, suggesting that both learning efficiency and LABJ tasks tapped into a similar underlying capacity to learn L2 vocabulary. However, the resulting measures are distinct, as the former assessment measured the capacity in a more sensitive manner, as it was embedded in the computer-assisted vocabulary retrieval practice. Possibly, efficient learners were more adept at using effective vocabulary learning strategies, such as elaboration with visualization or the keyword technique (Nation, 2022). Although Zerr *et al.* (2018) did not find a connection between strategy use and learning efficiency, further research could shed light on the intricate interplay among learning efficiency, aptitude, and strategy.

To further examine the role of different vocabulary learning schedules in moderating the predictive power of learning efficiency and aptitude (RQ2), a series of multiple regression analyses were conducted for each practice schedule group (i.e. separately for the Short–Short, Short–Long, Long–Short, and Long–Long datasets). As shown in Figure 1, consistent with the full-sample analysis, the learning-efficiency score in Session 2 was consistently the strongest predictor of vocabulary knowledge retention. However, while the learning-efficiency score in Session 1 failed to predict vocabulary retention in the full-sample analysis, it accounted for 9%–20% of the variance in the delayed posttest scores achieved by individual groups. This finding suggests that the learning-efficiency score becomes more sensitive when measured within specific learning conditions, such as those involving short- or long-spaced schedules. More specifically, in the productive posttest, the learning-efficiency score at Session 1 was particularly prominent in the Short–Long group, as supported by the generalized linear mixed-effects modeling. The reason why

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

this score at Session 1 was more influential in predicting posttest outcomes in the Short–Long group, however, remains unclear. Future research should explore how the relative importance of the learning-efficiency score varies at different points of (re)learning sessions.

Another observation relates to the LABJ score, which was the second strongest predictor – almost reaching statistical significance – albeit of the receptive test score for the Long–Long group only. Given that this group was exposed to the most demanding/difficult learning conditions due to the longer spacing between target lexical items in both sessions (Suzuki *et al*., 2019), aptitude could have played a more prominent role in this case. This observation is consistent with the general ATI pattern, suggesting that aptitude gains in importance when greater learning demands are imposed on learners (DeKeyser, 2019). Furthermore, the *receptive* test format of the LABJ mirrored that of the delayed receptive test, which could have amplified the role of aptitude in this receptive posttest. This nuanced pattern suggests that even subtle differences in learning modality (e.g. L2-to-L1 retrieval versus L1-to-L2 retrieval) can influence the predictive power of aptitude. Accordingly, it is possible that the learning-efficiency score emerged as such a strong predictor of vocabulary retention because it captured the ability to memorize vocabulary in the same format as the productive delayed posttest (i.e. L1-to-L2 retrieval). Indeed, as shown in Table 5, the relative importance of the learning-efficiency score in Session 2 was higher for the productive posttest than for the receptive posttest performance in the remaining three groups (Short–Short, Short–Long, and Long–Short) that had less demanding learning schedules.

## 7. Limitations and suggestions for further research

While the current study provides valuable insights into the relationship among learning efficiency, aptitude, and vocabulary retention, it was subject to several limitations that warrant consideration. First, the current experiment focused exclusively on the learning of written words. Acquisition of auditory lexical knowledge has been recognized as important for establishing robust written and auditory lexical knowledge for successful L2 skills, not just for reading and writing but also for listening and speaking (e.g. Saito, Uchihara, Takizawa & Suzukida, 2023; Uchihara, Webb, Saito & Trofimovich, 2022). Additionally, considering the proficiency level of the participants, who were non-advanced L2 learners (CEFR A2/B1), the lexical items used in the experiment were of low frequency and outside their current level of proficiency. Consequently, it is plausible that a significant number of participants struggled to assimilate the newly introduced vocabulary into their existing mental lexicon, a crucial aspect of vocabulary acquisition. This limitation is due to the inherent drawbacks of flashcard learning, which typically focuses on initial form–meaning mapping and may not necessarily facilitate the learning of other dimensions such as contextual vocabulary use (Nakata, 2020). Future research should extend to include vocabulary learning with more appropriate lexical levels as well as implementing different learning formats with different modality to explore the generalizability of current findings.

Second, the experimental design adopted in the original study that yielded data for reanalysis here consisted of only two learning sessions, limiting our understanding of the long-term implications of these predictors. Therefore, this shortcoming could be mitigated by adopting a longitudinal design in future research. For example, by gathering data over an extended period, such as one academic semester, researchers would be able to examine how learning efficiency and aptitude interact with vocabulary retention over time.

Third, although the LABJ is an established instrument for measuring language aptitude among Japanese learners (Sasaki, 1993) and demonstrated high reliability in this study, a close inspection of the dataset revealed that the distribution was negatively skewed (skewness $= -1.40$). Although the skewness score is within an acceptable range (± 2, according to Field, 2009), it suggests that the test may not have been sufficiently challenging for many participants, potentially affecting the validity of our findings. Therefore, the authors of future studies in this domain could increase the

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

difficulty of this subtest (e.g. by expanding the number of words in the learning list) to more accurately gauge the range of participants' aptitudes and its role in vocabulary learning and retention.

Finally, ATI patterns in this study need to be re-evaluated in future (replication) research. Because ATI patterns involve the interaction effect between treatment and individual difference factors, it is more complex and hence difficult to replicate the same pattern than simple effects of intervention itself (for discussion, see DeKeyser, 2019; Granena & Yilmaz, 2018).

## 8. Pedagogical and theoretical implications

From a pedagogical standpoint, the findings yielded by this study have important implications for optimizing CALL systems. For instance, the criterion for learning could be adjusted based on an individual's learning efficiency, thus tailoring the amount of vocabulary practice required. This is especially relevant in light of a recent review of flashcard apps for smartphones, which revealed that in the majority of such apps, target items are removed from the practice list after eliciting the correct response a fixed number of times (Kaitsu, 2023). Because less efficient learners tend to forget a greater number of lexical items they have been exposed to, they are more likely to benefit from additional vocabulary practice with more stringent criteria (e.g. requiring multiple successful retrievals before lexical items are excluded from further practice) to maximize their learning performance (Karpicke & Roediger, 2008). In contrast, because more efficient learners are more likely to retain their lexical knowledge, target items may be removed from practice after reaching a less strict criterion, such as one successful retrieval. This kind of personalized approach can make computer-assisted vocabulary learning more cost effective.

Theoretically, the findings obtained in the current study suggest that the traditional measurement of rote memory ability in language aptitude tests may be a less useful predictor in a specific learning context such as computer-assisted intentional vocabulary learning. In the CALL environment, where the information about learners' ability can be collected during the actual learning process, the concept of aptitude can be more flexibly reconfigured to enhance its predictive utility for L2 learning outcomes and to inform the individualization of L2 practice. Traditionally, aptitude has been considered a relatively stable trait or ability to learn an L2 efficiently (Carroll, 1981; Doughty, 2019). However, the learning-efficiency score that can be derived from the CALL system can serve as a more flexible and sensitive measure capable of dynamically capturing learners' abilities in a variety of contexts (for a complex dynamic systems view of individual differences, see Lowie & Verspoor, 2019). Most importantly, with the help of CALL systems, aptitude can be measured and updated in situ rather than being assessed only once prior to the learning activity.

## 9. Conclusions

The objective of this study was to evaluate the utility of learning efficiency and language aptitude as predictive variables in computer-assisted intentional vocabulary practice. The findings presented in this work underscore the importance of learning efficiency (i.e. the number of trials required to reach a learning criterion) as a predictor of long-term vocabulary retention in the CALL environment. Traditional language aptitude measures, conversely, demonstrated limited predictive utility, suggesting the need for the re-evaluation of their role in computer-assisted vocabulary learning personalization. The influence of learning schedules on the predictive strength of learning efficiency supports the benefit of context-sensitive assessments, such as those performed within different spacing schedules. In summary, these findings could inform the individualization of vocabulary practice in CALL systems, allowing for dynamic adjustments based on learners' actual performance metrics.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

**Supplementary material.** To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S095834402400020X

**Data.** The R script and the raw data file are available in https://osf.io/gzamc/?view_only=164f80040ef74e8697adb61748742bbe

**Acknowledgements.** This research was supported by JSPS KAKENHI grant (23K20482), awarded to the first author, and JSPS KAKENHI grant (22K00743), awarded to the second author. We greatly appreciate the invaluable suggestions given by the editor and anonymous reviewers. We would also like to thank Dr Keiko Hanzawa for her help with data collection.

**Ethical statement and competing interests.** All respondents participated in the study voluntarily and provided informed consent prior to the commencement of the study, in accordance with the research practices involving human participants. Throughout the study and the subsequent analyses, we ensured the data confidentiality and anonymity of all participants. No conflicts of interest are present in relation to this research. The authors declare no use of generative AI.

## References

Bjork, R. A. (1994) Memory and metamemory considerations in the training of human beings. In Metcalfe, J. & Shimamura, A. P. (eds.), *Metacognition: Knowing about knowing*. Cambridge: The MIT Press, 185–205. https://doi.org/10.7551/mitpress/4561.003.0011

Carroll, J. B. (1981) Twenty-five years of research on foreign language aptitude. In Diller, K. C. (ed.), *Individual differences and universals in language learning aptitude*. Rowley: Newbury House, 83–118.

Carroll, J. B. & Sapon, S. M. (1959) *Modern Language Aptitude Test: MLAT*. New York: Psychological Corporation.

Chalmers, J., Eisenchlas, S. A., Munro, A. & Schalley, A. C. (2021) Sixty years of second language aptitude research: A systematic quantitative literature review. *Language and Linguistics Compass*, 15(11): Article e12440. https://doi.org/10.1111/lnc3.12440

Dahlen, K. & Caldwell-Harris, C. (2013) Rehearsal and aptitude in foreign vocabulary learning. *The Modern Language Journal*, 97(4): 902–916. https://doi.org/10.1111/j.1540-4781.2013.12045.x

DeKeyser, R. M. (2019) Aptitude-treatment interaction in second language learning: Introduction to the special issue. *Journal of Second Language Studies*, 2(2): 165–168. https://doi.org/10.1075/jsls.00007.int

Doughty, C. J. (2019) Cognitive language aptitude. *Language Learning*, 69(S1): 101–126. https://doi.org/10.1111/lang.12322

Field, A. P. (2009) *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks: SAGE Publications.

Granena, G. & Yilmaz, Y. (2018) Aptitude-treatment interaction in L2 learning: A research synthesis. *Studies in English Education*, 23(4): 803–830. https://doi.org/10.22275/SEE.23.4.02

Kaitsu, T. (2023) *Analysis of flashcard apps for second language vocabulary learning*. Rikkyo University, unpublished master's thesis.

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

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

Larsen-Freeman, D. & Cameron, L. (2008) *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Lee, H., Warschauer, M. & Lee, J. H. (2019) Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL*, 31(2): 135–149. https://doi.org/10.1017/S0958344018000162

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

Li, S. (2016) The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, 38(4): 801–842. https://doi.org/10.1017/S027226311500042X

Li, Y. & Hafner, C. A. (2022) Mobile-assisted vocabulary learning: Investigating receptive and productive vocabulary knowledge of Chinese EFL learners. *ReCALL*, 34(1): 66–80. https://doi.org/10.1017/S0958344021000161

Lowie, W. M. & Verspoor, M. H. (2019) Individual differences and the ergodicity problem. *Language Learning*, 69(S1): 184–206. https://doi.org/10.1111/lang.12324

Malone, J. (2018) Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, 40(3): 651–675. https://doi.org/10.1017/S0272263117000341

Mizumoto, A. (2023) Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, 73(1): 161–196. https://doi.org/10.1111/lang.12518

Muñoz, C., Pattemore, A. & Avello, D. (2022) Exploring repeated captioning viewing as a way to promote vocabulary learning: Time lag between repetitions and learner factors. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2113898

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

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/10.1017/S0272263116000280

Nakata, T. (2020) Learning words with flash cards and word cards. In Webb, S. (ed.), *The Routledge handbook of vocabulary studies*. New York: Routledge, 304–319. https://doi.org/10.4324/9780429291586-20

Nakata, T., Suzuki, Y. & He, X. (2023) 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

Nakata, T. & Webb, S. (2016) Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, 38(3): 523–552. https://doi.org/10.1017/S0272263115000236

Nation, I. S. P. (2006) How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1): 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2013) Commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, 2(1): 32–38. https://doi.org/10.7820/vli.v02.1.nation

Nation, I. S. P. (2022) *Learning vocabulary in another language* (3rd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009093873

Pawlak, M. & Kruk, M. (2022) *Individual differences in computer assisted language learning research*. New York: Routledge. https://doi.org/10.4324/9781003240051

Ruiz, S., Rebuschat, P. & Meurers, D. (2021) The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research*, 25(4): 510–539. https://doi.org/10.1177/1362168819872859

Ruiz, S., Rebuschat, P. & Meurers, D. (2023) Supporting individualized practice through intelligent CALL. In Suzuki, Y. (ed.), *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology*. New York: Routledge, 119–143. https://doi.org/10.4324/9781003414643-7

Saito, K., Uchihara, T., Takizawa, K. & Suzukida, Y. (2023) Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S027226312300044X

Sasaki, M. (1993) Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning*, 43(3): 313–344. https://doi.org/10.1111/j.1467-1770.1993.tb00617.x

Seibert Hanson, A. E. & Brown, C. M. (2020) Enhancing L2 learning through a mobile assisted spaced-repetition tool: An effective but bitter pill? *Computer Assisted Language Learning*, 33(1–2): 133–155. https://doi.org/10.1080/09588221.2018.1552975

Serfaty, J. & Serrano, R. (2024) Practice makes perfect, but how much is necessary? The role of relearning in second language grammar acquisition. *Language Learning*, 74(1): 218–248. https://doi.org/10.1111/lang.12585

Stockwell, G. (2022) *Mobile assisted language learning: Concepts, contexts and challenges*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108652087

Suárez, M. M. & Gesa, F. (2019) Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference? *The Language Learning Journal*, 47(4): 497–517. https://doi.org/10.1080/09571736.2019.1617768

Suzuki, Y. (2019) Individualization of practice distribution in second language grammar learning: A 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

Suzuki, Y. & DeKeyser, R. (2017) Exploratory research on second language practice distribution: An aptitude x treatment interaction. *Applied Psycholinguistics*, 38(1): 27–56. https://doi.org/10.1017/S0142716416000084

Suzuki, Y., Nakata, T. & DeKeyser, R. (2019) The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, 103(3): 713–720. https://doi.org/10.1111/modl.12585

Uchihara, T., Webb, S., Saito, K. & Trofimovich, P. (2022) Does mode of input affect how second language learners create form–meaning connections and pronounce second language words? *The Modern Language Journal*, 106(2): 351–370. https://doi.org/10.1111/modl.12775

Webb, S. (2005) Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1): 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S., Yanagisawa, A. & Uchihara, T. (2020) How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, 104(4): 715–738. https://doi.org/10.1111/modl.12671

Wray, A. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press

Yu, A. & Trainin, G. (2022) A meta-analysis examining technology-assisted L2 vocabulary learning. *ReCALL*, 34(2): 235–252. https://doi.org/10.1017/S0958344021000239

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K. & McDermott, K. B. (2018) Learning efficiency: Identifying individual differences in learning rate and retention in healthy adults. *Psychological Science*, 29(9): 1436–1450. https://doi.org/10.1177/0956797618772540

## About the authors

**Yuichi Suzuki** is an associate professor at Waseda University. He is interested in bridging the gap between theory and practice in instructed second language acquisition research. His edited volume titled *Practice and Automatization in Second Language Research: Perspectives From Skill Acquisition Theory and Cognitive Psychology* was published by Routledge in 2023.

**Tatsuya Nakata** is a professor at the College of Intercultural Communication, Rikkyo University, Japan. His research interests include second language vocabulary acquisition and computer-assisted language learning. He teaches courses in second language acquisition, as well as supervising postgraduate research.

**Xuehong (Stella) He** is a lecturer in applied linguistics at Swansea University, UK. Her research interest is instructed second language acquisition, with special attention to vocabulary teaching and learning. She mainly adopts cognitive and psycholinguistic approaches, including eye-tracking technology.

https://doi.org/10.1017/S095834402400020X Published online by Cambridge University Press