*Studies in Second Language Acquisition* (2024), **46**, 663–685
doi:10.1017/S0272263124000184

RESEARCH ARTICLE

# Effects of retrieval schedules on the acquisition of explicit, automatized-explicit, and implicit knowledge of L2 collocations

Nan Fang[1], Irina Elgort[2] and Zhuo Chen[3]

[1]School of Foreign Studies, Shaoguan University, Shaoguan, China; [2]Centre for Academic Development and School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington, New Zealand; [3]Department of General Courses, Guangzhou Panyu Polytechnic, Guangzhou, China
**Corresponding author:** Nan Fang; Email: fnfangnan@sgu.edu.cn

(Received 21 November 2022; Revised 27 December 2023; Accepted 26 February 2024)

## Abstract

This study investigates the effects of retrieval schedules on the acquisition of second language (L2) collocations. Chinese learners of English first studied 36 target verb-noun collocations using flashcards and form-meaning matching practice. Subsequently, the participants practiced retrieving the target collocations from memory, following either a massed (consecutive) or spaced schedule. After each retrieval attempt, corrective feedback was provided. The acquisition of L2 collocations was measured by near-immediate and 1-week delayed posttests that assessed explicit knowledge with an offline form recall task, automatized explicit knowledge using an online acceptability judgment task, and implicit knowledge with an online collocation priming (lexical decision) task. Results showed equal learning effects of massed and spaced retrieval at both posttests of explicit knowledge and the near-immediate posttest of automatized explicit knowledge. The spacing effect was observed for the implicit knowledge across the two posttests and the automatized explicit knowledge at the delayed posttest.

## Introduction

Multiword expressions (MWEs, e.g., collocations, *kick a ball*; idioms, *kick the bucket*; phrasal verbs, *turn down*) are essential in developing second language (L2) vocabulary mastery (Schmitt, 1998), proficiency (Howarth, 1998), and fluency (Wray, 2002). They help language speakers effectively and efficiently communicate in real time. The knowledge of MWEs is multifaceted; it goes beyond being able to recognize an MWE and explain its meaning. A key benefit of MWE mastery is speakers' ability to process these expressions fluently, in real-time language use (Siyanova-Chanturia & Van Lancker Sidtis, 2019).

Learning MWEs is challenging for L2 learners, especially if their exposure to the target language is limited. Therefore, investigating the relative effectiveness of different approaches to learning MWE is a worthwhile second language acquisition (SLA)

research topic. Strong and Boers (2019a, 2019b), for instance, found that *retrieval practice* (i.e., opportunities for explicit retrieval of previously studied information from memory, as a learning event) is an effective approach to learning one type of L2 MWEs —phrasal verbs. This finding supports the theoretical claim that successfully retrieving recently encoded information from memory benefits its retention (known as the testing or retrieval effect; Roediger & Karpicke, 2006). This advantage of retrieval over restudying reported in learning and retention research (Roediger & Karpicke, 2011) may be understood as a function of processing difficulty (i.e., the greater the effort at encoding, the greater the retention; Bjork, 1994), and/or levels of processing during retrieval, with deeper processing (such as semantically rich processing) resulting in better knowledge retention than shallow processing (Craik & Lockhart, 1972; Craik and Tulving, 1975; Hintzman, 1976).

Furthermore, *retrieval schedules* (i.e., retrieval episodes distributed consecutively, *massed*, or further apart, *spaced*) affect learning and retention. Optimal retrieval difficulty (i.e., *desirable difficulties*) can maximize knowledge retention and transfer (Bjork, 1994; Schmidt & Bjork, 1992; for a recent discussion in the L2 field, see Suzuki et al., 2019a, 2019b). The testing/retrieval effect is predicted to be greater when retrieval is more effortful and requires deeper processing (see Roediger & Karpicke, 2011, for a review), for example, when retrieval involves meaning elaboration and retrieval episodes are spaced rather than massed (e.g., Balota et al., 2006; Karpicke & Roediger, 2007). When retrieval episodes are accompanied by feedback, the feedback information is likely to receive more attention and be processed deeper in the spaced than the massed condition because it is perceived as less familiar.

### Measures of MWE knowledge and processing

When evaluating MWE instructional and learning approaches, we need to look into what it means to "know" an MWE. As argued above, one of the main advantages of MWE mastery is the ease and automaticity of real-time language processing and use (Siyanova-Chanturia & Van Lancker Sidtis, 2019). Therefore, besides commonly used offline posttests of form and meaning recall, SLA researchers can also use online (real-time) processing measures to test the fluency and automaticity of MWE processing under time pressure (Sonbul & Schmitt, 2013; Toomer & Elgort, 2019). Notably, the effect of levels of processing may vary for explicit and implicit knowledge and processing measures (Hamann, 1990; Newell & Andrews, 2004; Roediger et al., 1989), suggesting that different levels of processing associated with massed and spaced *retrieval schedules* may not hold the same benefits for the development of and access to different types of knowledge (Ullman & Lovelett, 2018).

Explicit knowledge is conscious knowledge *about* something, such as facts, meanings, and experiences (e.g., the expression "*shake hands*" signifies "*agreement*"); it can be gained very quickly from a single learning episode. This knowledge can be automatized via repeated exposure and use (e.g., readers are likely to judge "*shake hands*" as a more acceptable phrase than "*shake hair*" under time pressure). Implicit knowledge, on the other hand, is gained gradually through repeated exposure, practice, and experience, often without explicit awareness (e.g., the word "*hands*" may be recognized and processed faster in reading if it follows its collocate "*shake*" than semantically unrelated/noncollocate "*swing*"); it is thus particularly important in fluent comprehension and production (Isbell & Rogers, 2021; Suzuki & DeKeyser, 2017).

L2 vocabulary learning studies suggest that spaced practice (including spaced retrieval practice) tends to be more effective than massed practice in acquiring explicit

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

knowledge of MWEs (for intentional learning of L2 collocations, for example, see Macis et al., 2021, and Yamagata et al., 2023); however, this may not necessarily be the case for the development of implicit knowledge (e.g., in their contextual word learning study, Nakata & Elgort, 2021, did not observe the spacing effect in the semantic priming task used as a proxy for tacit knowledge). Therefore, more research is needed to examine how spacing (i.e., the distribution of repetitions over time) affects the development of different types of MWE knowledge and access to this knowledge in real-time processing. Perhaps combining different learning and memory enhancement techniques (namely, spacing and retrieval) may facilitate not only the development of explicit and automatized explicit knowledge of MWEs but also their implicit knowledge and real-time processing.

In the present study, therefore, we investigate whether the spacing effect in retrieval practice of L2 collocations (defined here as the distribution of retrieval episodes over trials) is observed in the outcome measures representing different types of collocational knowledge. Although L2 collocation learning research has begun to examine the development of implicit knowledge (operationalized as collocational priming in lexical decisions; Sonbul & Schmitt, 2013; Toomer & Elgort, 2019), these studies have not tested whether retrieval spacing affects the acquisition of implicit knowledge.

In grammar research (e.g., Suzuki, 2017), implicit knowledge has been distinguished from automatized explicit knowledge. The latter is commonly measured using a timed sentence grammaticality judgment task. The knowledge measured is considered explicit because the task instructions raise awareness of the linguistic knowledge being measured; the knowledge is considered automatized explicit because it is accessed under time pressure. In L2 collocational processing research, a parallel task is a timed acceptability judgment task, in which participants judge whether a word combination is acceptable or not in a target language under time pressure (e.g., Öksüz et al., 2021; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010). Recently, this task has been used to measure the acquisition of automatized explicit knowledge of L2 MWEs (Jeong & DeKeyser, 2023; Northbrook et al., 2022).

Measuring the development of different aspects of knowledge helps us gain a more nuanced understanding of the effects of instructional techniques and learning approaches (Schmitt, 2022). Although retrieval practice appears to be effective in gaining explicit knowledge of L2 MWEs (Strong & Boers, 2019a, 2019b), further studies are needed to optimize the use of this practice format for developing different knowledge types. Specifically, our study goes beyond traditional posttests of explicit knowledge but also investigates how massed and spaced retrieval schedules affect the development of automatized explicit and implicit knowledge of L2 collocations.

### Retrieval practice and L2 MWE learning

Research evaluating MWE exercises in published English as foreign language textbooks found certain trial-and-error practice exercises problematic. In these practice exercises, learners first complete cloze or multiple-choice tasks with MWEs they need to learn and then receive corrective feedback. Learners whose responses are incorrect risk encoding undesirable associations (e.g., encoding, **talk volumes* instead of *speak volumes*; Boers et al., 2017). The corrective feedback in such exercises may not be sufficient to reverse these erroneous associations (Stengers & Boers, 2015; Strong & Boers, 2019b). On the other hand, tasks that necessitate retrieval practice following initial learning (also common in L2 textbooks) do not cause such problems, presumably because learners

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

are exposed to intact MWEs first, which allows them to form memory traces for the correct word combinations. The advantage of retrieval practice over trial-and-error exercises in deliberate L2 MWE learning was observed in immediate and delayed posttests (Strong & Boers, 2019a, 2019b).

Strong and Boers (2019b) operationalized trial-and-error practice as gap-fill tasks (e.g., *hang* ____—*spend time with friends*) followed by corrective feedback; the retrieval practice was operationalized as a form-meaning association procedure, where gap-fill tasks with corrective feedback were preceded by intact MWE presentation. The researchers argued that in deliberate learning and practice of L2 MWEs, learning procedures that result in fewer erroneous associations between the component words of MWEs lead to more accurate knowledge. These studies show that presenting intact L2 MWEs to learners upfront affords the creation of stronger associations between the component words of the MWEs, and retrieving these correct associations from memory improves short- and long-term knowledge retention (Karpicke & Roediger, 2007).

However, the benefits of retrieval practice were observed by Strong and Boers in a study with L2 phrasal verbs, and it is unclear whether these findings are generalizable to other MWE types, as different MWEs present different challenges to L2 learners. English phrasal verbs that usually comprise high-frequency words are challenging due to their semantic opaqueness and polysemy (Garnier & Schmitt, 2015), while the main difficulties in learning L2 (English) verb collocations are associated with the choice of verbs, due to the interference from learners' knowledge of corresponding first language (L1) collocations (Laufer & Waldman, 2011; Nesselhauf, 2003). Further research is needed, therefore, to test whether retrieval practice benefits the learning and acquisition of L2 verb collocates (Boers et al., 2014; Szudarski & Carter, 2016; Tsai, 2020; Webb & Kagimoto, 2009).

### *Spacing and L2 MWE learning*

Most L2 vocabulary spacing studies examined the learning of single words (e.g., Nakata & Suzuki, 2019); so far, the spacing effect in MWE learning has been addressed in only a handful of studies. In a classroom study with Iranian junior high school students, Farvardin (2019) observed the spacing effect in intentional learning of L2 English collocations (assessed with near-immediate and 2-week/4-week delayed posttests of form recognition and meaning recall).

Macis et al. (2021) investigated the effect of spacing in incidental and intentional learning of L2 collocations. In a between-participants design, two groups of Arabic speakers learned 25 English adjective-noun collocations incidentally (through reading) or intentionally (memorizing and studying target collocations embedded in concordance lines). Two spacing schedules of a whole learning event (study plus retrieval) were used; in the massed condition, five collocations were repeated five times per session; in the spaced condition, each of the 25 collocations was presented once per session. The spacing effect was observed on a 3-week delayed cued form-recall posttest in the intentional learning group.

In Yamagata et al. (2023), Japanese high school students engaged in the form-focused practice of English verb-noun collocations in a classroom setting. The collocations were practiced seven times per session, following three distribution schedules over 3 weeks (three sessions each week): node-massed (i.e., massed repetitions of the same nodes each day, e.g., Week 1, *run a fever*/*story*/*finger*), collocation-massed

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

(i.e., massed repetitions of collocations of different nodes each day and spaced repetitions of collocations of the same nodes across 3 weeks, e.g., Week 1, *run a fever*; Week 2, *run a story*; Week 3, *run a finger*), and collocation-spaced (i.e., massed repetition of the same nodes each day and spaced repetitions of individual collocations of the same node across 3 weeks, e.g., Week 1/2/3, *run a fever*/*story*/*finger*). Their learning treatment included retrieval attempts of a given target verb (e.g., *run*) in/out of context, the target collocation in context, and other learning procedures (i.e., presentation, translation, and quizzes). The collocation-spaced schedule group (i.e., the condition requiring spaced retrieval of individual collocations) outperformed the massed groups in near-immediate and delayed collocation/verb gap-filling posttests. This finding seems to align with the advantage of spaced retrieval over massed retrieval found in L2 single-word learning studies (Karatas et al., 2021; Koval, 2022). However, like Macis et al. (2021), Yamagata and colleagues (2023) focused on the distribution of the complete learning event, including study and retrieval, possibly conflating the effects of study spacing and retrieval spacing. Importantly, both studies only tested offline explicit knowledge of collocations. In summary, little is known about the effects of retrieval schedules on the acquisition of different aspects of L2 collocational knowledge.

### *Present study*

We investigate the effect of retrieval schedules on the acquisition of implicit, automatized explicit, and explicit knowledge of L2 collocations. Following the findings of Strong and Boers (2019b), study participants first studied collocations using flashcards and decontextualized form-meaning matching tests (i.e., familiarization stage), after which they engaged in either consecutive (massed) or distributed (spaced) retrieval practice (i.e., retrieval stage). Near-immediate and announced 1-week delayed posttests were administered to capture the initial learning and retention of collocational knowledge, respectively (Soderstrom & Bjork, 2015). The participants' explicit knowledge of the target L2 collocations was measured using a form recall task (Sonbul & Schmitt, 2013; Toomer & Elgort, 2019), their automatized explicit knowledge was measured using an online acceptability judgment task (Jeong & DeKeyser, 2023; Northbrook et al., 2022), and their implicit knowledge was measured using a primed lexical decision task (Sonbul & Schmitt, 2013; Toomer & Elgort, 2019). The following research questions guided the study:

1. Does retrieval schedule affect the acquisition of explicit knowledge of L2 collocations? If yes, how?
2. Does retrieval schedule affect the acquisition of automatized explicit knowledge of L2 collocations? If yes, how?
3. Does retrieval schedule affect the acquisition of implicit knowledge of L2 collocations? If yes, how?

## Methods

### *Participants*

Twenty-nine undergraduate students (28 women), English majors, from an intact class participated in the study. Each participant received 50 CNY and a small gift. Their age ranged from 18 to 21 years ($M$ = 19.31; $SD$ = .66). All participants have learned English through formal education, with Mandarin Chinese as a medium of instruction. The

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

mean starting age of learning English was 9.07 years (*SD* = 2.39), and the mean length (in years) of learning was 10.24 (*SD* = 2.28). None of the participants had visited an English-speaking country. Their English vocabulary knowledge was estimated with LexTALE (Lemhöfer & Broersma, 2012), and the mean score was 52.93% (*SD* = 8.83%), which suggests an intermediate proficiency level.

## Materials

### Collocations

The collocations used in our study (see Appendix S1) were either adjacent (e.g., *lend weight*) or nonadjacent with a determiner (e.g., *a*, *an*, *the*) between the verb and the noun (e.g., *attend a clinic*). They were developed using the Academic Subcorpus of the British National Corpus (BNC Consortium, 2007), or the BNC-AC. The chosen collocations met the set thresholds (collocational frequency >10, *t*-scores >2, and MI >3) based on the enTenTen20 corpus (English Web Corpus 2020; Jakubíček et al., 2013), and the constituent words in the collocations were four to eight letters long. The first author, a native Chinese speaker with advanced English proficiency, selected a pool of potential incongruent collocations based on the intuitive judgment of L1-L2 incongruency. These collocations were randomly listed in two translation tests and given to 15 native Chinese speakers with high English proficiency (graduate students in teaching English to speakers of other languages and applied linguistics). They were instructed to translate the English words into as many Chinese translations as possible. The most translated words were identified as dominant translation equivalents. In addition, the English-Chinese version of *Oxford Collocations Dictionary* (McIntosh, 2015) was consulted for the Chinese translation of these English collocations. Further, the collocation renderings were compared against their word-for-word renderings to confirm that these collocations were incongruent. The selected English collocations were administered to 30 English learners from the same population as that used in this study in a productive L1-to-L2 translation test (e.g., Laufer & Girsai, 2008; Webb & Kagimoto, 2011). Based on the translation test results, collocations were divided into two groups: target collocation (score <10% accuracy) and baseline familiar collocations (score >80% accuracy).

### Other stimuli

In the two online posttests (acceptability judgment and primed lexical decision), the stimuli included target collocations, familiar collocations, and their matched controls (i.e., nonce phrases consisting of a noncollocate and the target word of a given collocation, e.g., *serve-notice* versus *compete-notice*), collocational and nonce-collocation fillers, and nonwords. The stimuli other than the target and familiar collocations were developed as follows. First, potential collocational fillers were selected from the BNC-AC and then searched in the enTenTen20 corpus to obtain lexical frequencies. The final collocational fillers had an enTenTen20-based MI and *t*-score higher than three and two, respectively. Second, other words in the nonce collocations (controls or fillers) were randomly selected from the 3,000-word families in English based on the BNC / Corpus of Contemporary American English word list (Nation, 2012). The potential nonce collocations were only kept if the component words did not commonly co-occur in the corpus (MI <3, *t*-score <2). Third, the nonword stimuli were created using the Wuggy software (Keuleers & Brysbaert, 2010). The final stimuli were checked to contain no duplicate words, and the items were four to eight letters long.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

Counterbalanced item lists were developed for the two online posttests. Each list included 18 target collocations, 18 nonce-collocation controls, 11 familiar collocations, and 11 nonce-collocation controls). In addition, the acceptability judgment lists included 18 nonce-collocation fillers and 18 collocational fillers (e.g., *cause-trouble*); the lists used in the primed lexical decision posttest contained 39 nonce-collocation fillers (e.g., *draft-shame*) and 97 word-nonword pairs[1] (e.g., *handle-notave*). Appendix S2 presents sample test lists of both tasks.

### *Final stimuli*

The learning targets (target collocations) were 36 unfamiliar incongruent English verb-noun collocations (e.g., *attend a clinic*, *lend weight*). In addition, 22 familiar English collocations (e.g., *spend a holiday*, *keep track*) were selected to establish the baseline for the online processing of L2 collocations. The familiar collocations were divided into two sets. One set (i.e., studied familiar collocation) was included in the familiarization stage (see below) and the posttests but not in the retrieval practice. The second set (i.e., unstudied familiar collocation) was only included in the posttests (but not in the learning treatment or the retrieval practice) and provided a baseline for two posttests (i.e., collocation priming and online acceptability judgments). This design allowed us to check whether exposure recency was a factor at the posttest stage.

### *Flashcards*

Flashcards (each containing a collocation, its definition, and an example sentence, all in English) were developed for the learning treatment. The English definitions were from online dictionaries (Cambridge, Collins, Merriam-Webster, Oxford, etc). The sentential materials were based on the enTenTen20 corpus (Jakubíček et al., 2013); the original concordances were minimally revised so that (a) they were meaningful and complete and (b) proper nouns (e.g., names) that may introduce comprehension difficulty were substituted or paraphrased (e.g., using pronouns and general terms). The sentences were further checked and minimally revised by a native speaker of English to ensure naturalness. The AntWordProfiler program (Anthony, 2014) was used to assess the lexical profiling of the English definitions and example sentences. The definitions were, on average, 8.67 words long ($SD$ = 2.97); 97.46% and 99.49% lexical coverage were reached, respectively, with the first 3,000 and 5,000 most frequent word families of English. The example sentences were, on average, 14.5 words long ($SD$ = 1.58); 95.9% and 99.01% lexical coverage were reached, respectively, with the first 3,000 and 5,000 most frequent word families. This was considered sufficient for the participants to understand the definitions and examples in the flashcards.

[1]The word-nonword ratio was .5 (i.e., an equal number of word and nonword trials). The relatedness proportion (i.e., the ratio of related trials to all word-word trials) in the primed lexical decision task was .3 (29/97; i.e., [18 target collocations + 18 matched nonce-collocations controls] + [11 familiar collocations + 11 matched nonce-collocations controls] + 39 nonce-collocation fillers). As one of the reviewers pointed out, the proportion of related trials in the primed lexical decision task (.3) was somewhat higher than .2, recommended for semantic priming experiments by McNamara (2005). However, our study complied with a key requirement of semantic priming, that is, using a short SOA (stimulus-onset asynchrony of 200 ms or less), as primes were displayed for only 150 ms in our experiment.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

### Learning treatment

The learning treatment consisted of familiarization and retrieval stages. In the familiarization stage, the participants studied target collocations and familiar collocations (hereafter "studied") using flashcards and decontextualized form-meaning matching tests. The flashcards ensured that each learning event was focused and discrete. Presenting intact collocations in the familiarization task provides an opportunity for the initial (baseline) encoding of the collocations' form-meaning mapping. This prior exposure is necessary for subsequent retrieval (Van den Broek et al., 2018) and for the development of automatized explicit knowledge, as proposed by skill acquisition theory (McLaughlin, 1987; Suzuki & DeKeyser, 2017). Furthermore, it can reduce the risk of forming erroneous lexical associations during retrieval (Boers et al., 2017; Strong & Boers, 2019b). The matching test was used to motivate the participants to genuinely study the form-meaning associations of the collocations. In the retrieval stage, the participants engaged in retrieval practice of the learning targets (but not the familiar collocations) and received corrective feedback.[2]

#### *Familiarization stage*

The familiarization stage of the learning treatment included the flashcard procedure (i.e., 47 collocations: 36 target collocations and 11 studied familiar collocation) and the form-meaning matching test. The flashcards were first presented one by one in sets of five or seven (Figure 1-a). Participants were instructed to study the association between each collocation and its definition and read the example sentence, presented on the same screen. A flashcard remained on the screen until participants pressed the space key. The flashcard activity was followed by a matching test on these five- or seven-pair sets (Figure 1-b). The participants were given as much time as needed to match the collocations and their definitions presented in separate columns and type in their responses. Regardless of the correctness of the response, a feedback screen (showing the test items, participants' responses, and correct answers, see Figure 1-c) was presented for a maximum of 100 s in the 5-pair sets and 140 s in the 7-pair sets (i.e., approximately 20 s per pair). The participants were instructed to check the answers; they could press the space key to terminate this display as soon as they finished reviewing the answers.

The purpose of the familiarization stage was to enable the participants to establish initial form-meaning associations for all target collocations before practicing their retrieval (Boers et al., 2017). Because some of the target collocations may have been partially familiar to some participants before the experiment, we allowed participants to complete the familiarization procedure at their own pace. For the purposes of retrieval practice, it was more important to confirm that all participants were at a similar level of accuracy in the form-meaning matching test—a proxy for similar levels of familiarity with the collocations prior to practicing their retrieval. The corrective feedback aimed to further reduce any differences in the participants' initial encoding of the target

[2]Because our goal for the familiarization task was to encode the target collocation and create their initial form—meaning associations—we did not control time-on-task for in the learning treatment. This is different from studies comparing the effect of different instructional/learning treatments, where time-on-task needs to be controlled. Furthermore, in our study, each retrieval attempt can be considered a discrete event that contributes to learning, regardless of the time needed to retrieve the missing word; the extra time taken in the retrieval task is unlikely to result in additional learning. We made some parts of the learning procedure (such as the feedback screen) self-paced to better align it with individual needs and to reduce the time of a rather long experiment, where possible.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

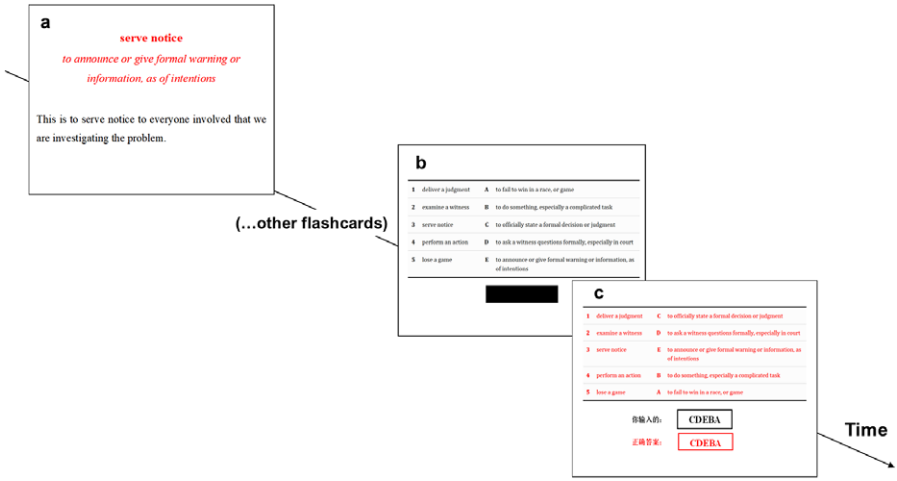


**Figure 1.** The familiarization stage procedure (1-a: a flashcard display; 1-b: a matching test display; 1-c: a feedback display).
*Note:* In Figure 1-c, the black and red box content display the responses and the correct answers, respectively.

collocations. The results of the familiarization stage confirmed that the participants could correctly match, on average, 95% of the target items (*SD* = 6%), and no participant did worse than 70% on this task. A practice session with five collocational flashcards and a five-pair matching test was conducted before the respective stages of the familiarization procedure.

### *Retrieval stage*

There were 72 trials in the retrieval stage of the learning treatment, namely, three retrieval episodes of 24 (of 36) target collocations. In a counterbalanced design, the target collocations were assigned to one of two retrieval conditions (i.e., R3_Massed or R3_Spaced, n = 12 each) or a baseline nonretrieval (R0) condition (n = 12). In the R3_Massed condition, the collocation was retrieved in three *consecutive* retrieval trials; in the R3_Spaced condition, the three retrieval trials for the same collocation were separated by 12 intervening retrieval trials of other collocations; in the R0 condition, the collocations were not included in the retrieval stage. Two practice trials (not including any familiar or target collocations) were presented before the retrieval stage.

For each trial, the screen displayed the following: (a) a collocation, in which the verb was replaced with an underlined space (e.g., _____ *notice*), (b) a definition, and (c) an input box, all displayed until a response was entered and submitted. The participants were instructed to fill in the missing verb for the collocations from the familiarization stage. After each retrieval, corrective feedback was presented for 10 s (i.e., allotting enough time to check the accuracy of their response briefly); participants could terminate the feedback in a self-paced manner. Different feedback screens were presented for correct and incorrect responses: for correct responses, the target collocation was presented in green with a tick next to it (Figure 2-$b_1$); for incorrect responses, the participant's response was presented in red with a cross next to it, and the target collocation was presented below in green with a tick (Figure 2-$b_2$). Thus, the correct collocation was always present in the feedback.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

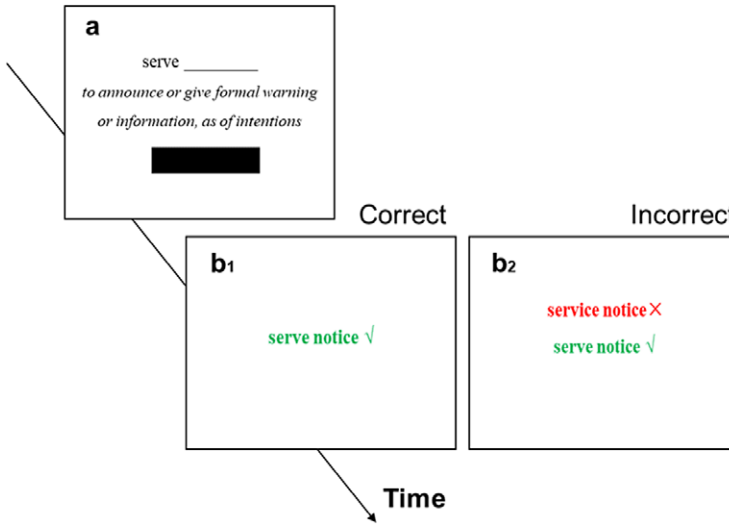


**Figure 2.** The retrieval practice procedure (2-a: a retrieval display; 2-$b_1$: a correct response display; 2-$b_2$: an incorrect response display).

## *Measures and posttests*

### *Form recall*

The form recall task measured participants' explicit knowledge of the 36 target collocations (Sonbul & Schmitt, 2013). We adopted the decontextualized format from Tsai (2020). The participants had to type the missing verb of the target collocations for which the initial letter was shown (e.g., *s____ notice*) to restrict responses to the target words as much as possible. The definition was also displayed. The participants were instructed to fill in the gap using the collocations from the learning treatment.

### *Acceptability judgment task*

An English acceptability judgment task was used to measure the automatized explicit knowledge of L2 collocations (Jeong & DeKeyser, 2023; Northbrook et al., 2022). The participants were instructed to decide whether a presented word sequence is an acceptable expression in English (Yamashita & Jiang, 2010, p. 657). In each trial, participants first saw a series of 12 asterisks in the middle of the screen, presented for 800 ms, followed by a 66-ms blank screen (Figure 3) and then a collocation or nonce-collocation phrase (e.g., *serve notice*; *compete notice*), which remained on the screen until response (or for the maximum of 4,000 ms). The task included ten practice trials (not including any two-word pairs from the experimental trials of the two online tasks) and 94 experimental trials (for details, see section, Materials, Other stimuli), in random order.

### *Primed lexical decision task*

An English primed lexical decision task was used to measure the implicit knowledge of L2 collocations (Sonbul & Schmitt, 2013). In this task, participants are presented with prime-target sequences and instructed to decide whether the target is an English word. The participants first saw a fixation (+) in the middle of the screen for 2,000 ms,

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

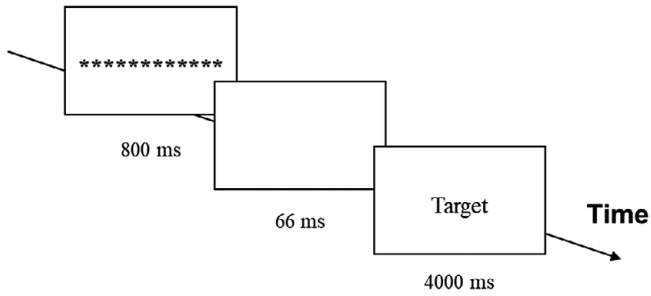


**Figure 3.** Experiment procedure of the acceptability judgment task.

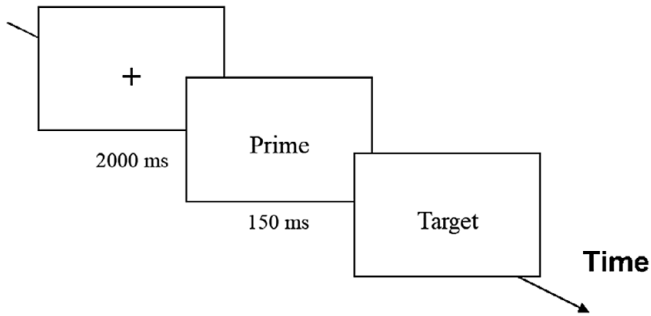


**Figure 4.** Experiment procedure of the primed lexical decision task.

followed by a 150-ms presentation of the prime, either the verb from the collocation or its substitution verb from the nonce collocation (e.g., *serve*/*compete*). The prime was immediately followed by the target (e.g., *notice*), which remained on the screen until a response was submitted (Figure 4). It included ten practice trials and 194 experimental trials (for details, see section, Materials, Other stimuli), in a random order. The determiner of the nonadjacent collocations was not included in the stimuli of this task. The presence of implicit knowledge is operationalized as collocation priming, namely, faster responses to the target (the terminal noun of the collocations, e.g., *notice*) preceded by its collocate verb prime (e.g., *serve*) compared with responses to the target preceded by a noncollocate verb prime (e.g., *compete*). The short prime duration and interstimulus interval were used to reduce participants' ability to develop and deploy explicit task strategies and facilitate the deployment of implicit knowledge. Because participants made lexical decisions only on the target (the last word of the collocation) and not on the prime, they were less likely to engage their explicit knowledge of the collocations. Thus, the collocation priming effect was hypothesized to primarily measure implicit knowledge of the collocations.

### *Self-reported knowledge tasks*

In a self-reported prior knowledge task (administered before the learning treatment), participants indicated whether the target collocations were known to them. In a self-reported additional exposure task (administered after the delayed posttests to account for any exposure to the target collocations between the near-immediate and delayed

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

posttests), participants reported whether they had encountered the target collocations during the interval. Both were yes/no tasks administered as an online questionnaire.

### Procedure

A day before the experiment, the participants completed LexTALE. The main study was conducted in a language lab, on an individual basis. It consisted of two parts, 1.5–2 hr in total. The first part began with the self-reported prior knowledge task, the learning treatment (familiarization stage followed by retrieval practice), the language background questionnaire (in Chinese, about 2–3 min long), and the near-immediate posttests (primed lexical decision, acceptability judgment, and form recall, in that order; the order of the tests was chosen to minimize the test-retest effect, with least explicit measures administered first). In the second part (1 week later), the participants returned for the announced delayed posttests and the additional self-reported exposure task. The posttests were completed in the same order in the near-immediate and delayed sessions; however, in the delayed primed lexical decision and acceptability judgment posttests, participants received alternative stimuli lists: if a participant saw an intact collocation in the immediate posttest, that participant saw a corresponding nonce collocation in the delayed posttest and vice versa. This design was used to counteract the potential test-retest effect.

## Data analysis

We preprocessed the data as follows. For the primed lexical decision, data of erroneous responses (18.67%) were excluded. Following Jiang (2012), we also excluded the extreme outliers with standardized residuals above 3 *SDs* and those with a response latency below 200 ms (9.63%). For the acceptability judgment, we excluded erroneous responses (36.06%) and responses with a response latency shorter than 450 ms (12.19%). The response times (RTs) were inverse-transformed (i.e., –1,000/RT) to reduce skewness in the distribution. The form recall responses were scored as 1 (correct) or 0 (incorrect) by the first author; spelling mistakes that did not interfere with the recognition of the intended answer were scored as correct responses. (e.g., **assump*, *assume*; 31 of 2088 responses, 1.48%) (e.g., Toomer & Elgort, 2019; Yamagata et al., 2023).

The data analysis consisted of a preliminary analysis and a primary analysis. The preliminary analysis examined the effect of Exposure in the familiarization stage on the subsequent online processing of the familiar collocations (studied familiar collocation/unstudied familiar collocation), for a more nuanced interpretation of the primary analysis findings. The primary analysis focused on the effect of Treatment (R0/R3_Massed/R3_Spaced). We fitted linear mixed-effects models to the RT data in the analysis of primed lexical decisions and acceptability judgments and generalized linear mixed-effects models to the binary response data in the analyses of form recall, using the lme4 R package (Bates et al., 2015). All initial models contained by-participants and by-items random intercepts, and random slopes for Test_time. The final models contained a random-effect structure supported by data (Matuschek et al., 2017). The alpha level was .05 in all analyses. The effect sizes (odds ratio [OR]; Cohen's *d* [*d*]) were interpreted following the general guidelines in Chen et al. (2010) (small OR = 1.68, medium OR = 3.47, large OR = 6.71) and Cohen (1988) (small *d* = .2; medium *d* = .5; large *d* = .8).

We began by fitting the most complex model and conducted backward stepwise model selection. The initial model of form recall included the following primary interest

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

predictors: Treatment (R0/R3_Massed/R3_Spaced), Test_time[3] (near-immediate/delayed posttest), and the Treatment × Test_time interaction. In the preliminary analyses for the acceptability judgment and primed lexical decision, the initial models included Exposure (nonce-collocation control/studied familiar collocation/unstudied familiar collocation), Test_time, and the Exposure × Test_time interaction. In the primary analyses, the initial models included Treatment (nonce-collocation control/R0/R3_Massed/R3_Spaced), Test_time, and the Treatment × Test_time interaction. The interactions were included because the effects of Treatment may differ for the near-immediate and delayed posttest (as shown in previous spacing and retrieval vocabulary learning studies). Post hoc analyses with the Tukey test were performed for significant interactions using the R package emmeans (Lenth et al., 2018).

The following covariates were also included in all the initial models: participants' L2 lexical proficiency (centered LexTALE scores) and whether the target included a determiner (absent/present). Furthermore, the initial models in the primary analyses included self-reported prior knowledge (yes/no) and additional exposure (yes/no) responses, as covariates. A back-stepping model simplification procedure resulted in the final models that contained the primary interest predictors; covariates and interactions were only kept when they improved the model fit. Additionally, we trimmed model residuals to 2.5 *SDs* after fitting the final RT models (Baayen et al., 2008).

## Results

Table 1 presents the descriptive statistics for the three near-immediate and delayed posttests by Treatment condition and Collocation type.

### *Results of form recall*

The form recall data of the target collocations showed reasonable reliability (near-immediate posttest: $\alpha = .874$; delayed posttest: $\alpha = .792$), according to the guidelines in Plonsky and Derrick (2016). The results (Table 2) showed significant effects of Treatment, Test_time, centered LexTALE scores, and self-reported additional exposure. When collapsed across the immediate and delayed posttests, the participants were more accurate in recalling target collocations in both the R3_Massed condition (OR = 3.39 [medium effect]) and the R3_Spaced condition (OR = 4.22 [medium effect]) than in the R0 condition. Although the R3_Spaced condition resulted in slightly higher mean accuracy than the R3_Massed condition (Table 1), the post hoc comparison results revealed that the difference was not statistically significant (R3_Massed versus R3_Spaced; $b = -.21$, OR = .81, $p = .305$ [small effect]). In all treatment conditions,

[3]One of the reviewers pointed out, and we agree, that near-immediate posttests could be considered an additional retrieval opportunity and may affect performance on the delayed posttests (e.g., Rogers, 2023). To some extent, we are able to account for the potential test-retest effect statistically by including Test_time (immediate/delayed) as a primary predictor in the models and by attempting to fit a by-participant and by-item random slopes in the random effects structure of the mixed-effect models. Importantly, any additional retrieval opportunity afforded by the near-immediate posttests was present for both spaced and massed practice condition. Finally, the participants received different stimuli lists in the immediate and delayed processing posttests (i.e., the intact collocations in the immediate posttest were replaced by nonce collocations in the delayed posttest, and vice versa). This design ensured that the intact target collocations were only presented once, either in the near-immediate or the delayed posttest for these tasks.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

**Table 1.** Descriptive results of the near-immediate and delayed posttests

| | | Form Recall | | Acceptability Judgment | | Primed Lexical Decision | |
|---|---|---|---|---|---|---|---|
| Condition | Type | Immediate Posttest | Delayed Posttest | Immediate Posttest | Delayed Posttest | Immediate Posttest | Delayed Posttest |
| Unstudied | Familiar collocation | — | — | 1077.56 (359.97) | 1118.16 (531.83) | 766.22 (331.59) | 648.3 (220.01) |
| | Control | — | — | 1433.21 (502.81) | 1319.41 (540.57) | 775.21 (373.84) | 699.35 (354.99) |
| Studied | Familiar collocation | — | — | 1083.38 (514.55) | 972.58 (378.66) | 687.6 (229.3) | 722.28 (343.45) |
| | Control | — | — | 1405.35 (446.24) | 1331.93 (399.11) | 723.57 (327.53) | 669.08 (269.49) |
| R0 | Target collocations | 26% (44%) | 12% (33%) | 1159.12 (429.36) | 1242.54 (518.12) | 813.68 (389.72) | 733.88 (276.24) |
| | Control | — | — | 1419.85 (468.46) | 1318.72 (465.66) | 838.02 (328.49) | 754.58 (262.83) |
| R3_Massed | Target collocations | 44% (50%) | 20% (40%) | 1117.31 (431.59) | 1226.83 (504.06) | 765.36 (335.08) | 711.47 (267.28) |
| | Control | — | — | 1477.81 (519.89) | 1354.72 (531.91) | 804.47 (386.47) | 741.92 (253.19) |
| R3_Spaced | Target collocations | 46% (50%) | 23% (42%) | 1119.73 (437) | 1180.18 (532.13) | 720.16 (315.25) | 733.04 (264.11) |
| | Control | — | — | 1439.53 (555.82) | 1271.21 (377.44) | 815.02 (391.37) | 717.51 (255.08) |

*Note*: The values are the mean accuracies for the form recall and response times (in milliseconds) for the acceptability judgment and primed lexical decision, with *SDs* in parentheses.

**Table 2.** Accuracy rates of form recall (target collocations): Fixed effects

| | *b* | 95% CI | *SE* | *z* | *OR* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | –1.95 | [–2.64, –1.27] | .35 | –5.6 | .14 | < .001 |
| Treatment = R3_Massed | 1.22 | [.89, 1.55] | .17 | 7.3 | 3.39 | < .001 |
| Treatment = R3_Spaced | 1.44 | [1.11, 1.77] | .17 | 8.57 | 4.22 | < .001 |
| Test_time = Delayed Posttest | –1.79 | [–2.33, –1.24] | .28 | –6.4 | .17 | < .001 |
| LexTALE (Centered) | .10 | [.04, .17] | .03 | 3.17 | 1.11 | .002 |
| Additional Exposure = Yes | .97 | [.6, 1.34] | .19 | 5.08 | 2.64 | < .001 |

*Note*: CI = confidence interval; Reference level: Treatment = R0; Test_time = near-immediate posttest; Additional exposure = no. Model formula: FR.accuracy ~ Treatment + Test_time + LexTALE_centered + Additional_exposure + (Test_time+1|Participant) + (LexTALE_centered +1|Target).

the participants recalled the target collocations more accurately at the immediate posttest than at the delayed posttest (OR = .17 [small effect]). In addition, the higher accuracy of form recall was associated with higher LexTALE scores (OR = 1.11 [small effect]) and self-reported additional exposures (OR = 2.64 [small effect]). The findings of form recall can be summarized as follows:

R3_Spaced = R3_Massed > R0

## *Results of acceptability judgments*

### *Preliminary analysis*

We tested whether the collocation processing advantage (i.e., faster processing of collocations compared with matched nonce-collocation controls) was observed for the familiar collocations. Results (see Table 3) showed this effect for both the unstudied familiar collocations ($d$ = .76 [medium effect]) and the studied familiar collocations ($d$ = 1.09 [large effect]), suggesting that the participants had automatized explicit knowledge of familiar L2 collocations that could be detected regardless of the exposure recency.

### *Primary analysis*

The Treatment × Test_time interaction was significant in the final model (see Table 4). The post hoc results on the Treatment × Test_time interaction are presented in Table 5. At the near-immediate posttest, the target collocations (R0: $d$ = .64; R3_Spaced: $d$ = .76 [medium effect]; R3_Massed: $d$ = .84 [large effect]) were judged faster than the nonce-collocation controls in all treatment conditions. The collocational advantage reported in the nonretrieval R0 condition suggested that any gain of automatized explicit knowledge observed for the target collocations in the retrieval practice treatment would likely be due to the cumulative effect of familiarization plus retrieval

**Table 3.** Response times of acceptability judgments (familiar collocations): Fixed effects

| | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | –.84 | [–.94, –.73] | .05 | –16.19 | –3.59 | < .001 |
| Exposure = Unstudied Familiar Collocation | –.18 | [–.22, –.13] | .02 | –7.69 | –.76 | < .001 |
| Exposure = Studied Familiar Collocation | –.25 | [–.3, –.21] | .02 | –11.27 | –1.09 | < .001 |
| Test_Time = Delayed Posttest | –.11 | [–.27, .04] | .08 | –1.49 | –.49 | .153 |

*Note*: CI = confidence interval; Reference level: Exposure = nonce-collocation control; Test_time = near-immediate posttest. Model formula: inverseRT ~ Exposure + Test_time + (Test_time +1|Participant) + (1|Target).

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

**Table 4.** Response times of acceptability judgments (target collocations): Fixed effects

| | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | –.85 | [–.94, –.76] | .04 | –19.27 | 3.20 | <. 001 |
| Treatment = R0 | –.17 | [–.23, –.11] | .03 | –5.21 | .64 | < .001 |
| Treatment = R3_Mass | –.22 | [–.28, –.16] | .03 | –7.24 | .84 | < .001 |
| Treatment = R3_Spaced | –.21 | [–.27, –.15] | .03 | –6.85 | .79 | < .001 |
| Test_time = Delayed Posttest | –.1 | [–.15, –.04] | .03 | –3.62 | .36 | < .001 |
| Determiner = Yes | .05 | [.02, .09] | .02 | 3.11 | .20 | .004 |
| Treatment = R0 × Test_time = Delayed | .13 | [.04, .22] | .05 | 2.87 | .50 | .004 |
| Treatment = R3_Mass × Test_time = Delayed | .20 | [.11, .29] | .05 | 4.40 | .75 | < .001 |
| Treatment = R3_Spaced × Test_time = Delayed | .11 | [.02, .20] | .05 | 2.46 | .42 | .014 |

*Note*: CI = confidence interval; Reference level: Treatment = nonce-collocation control; Test_time = near-immediate posttest; Determiner = no determiner. Model formula: inverseRT ~ Treatment * Test_time + Determiner + (1|Participant) + (1|Target).

**Table 5.** Post hoc comparisons of the response times for the Treatment × Test_Time interaction in the analysis of acceptability judgments (target collocations)

| | Contrast | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|---|
| Immediate Posttest | Control vs. R0 | .17 | [.39, .89] | .03 | 5.20 | .64 | < .001 |
| | Control vs. R3_Mass | .22 | [.60, 1.08] | .03 | 7.23 | .84 | < .001 |
| | Control vs. R3_Spaced | .21 | [.55, 1.02] | .03 | 6.84 | .79 | < .001 |
| | R0 vs. R3_Mass | .05 | [–.07, .47] | .04 | 1.47 | .20 | .454 |
| | R0 vs. R3_Spaced | .04 | [–.12, .42] | .04 | 1.09 | .15 | .696 |
| | R3_Mass vs. R3_Spaced | –.01 | [–.31, .21] | .03 | –.41 | .05 | .976 |
| Delayed Posttest | Control vs. R0 | .04 | [–.12, .39] | .03 | 1.09 | .14 | .698 |
| | Control vs. R3_Mass | .03 | [–.16, .35] | .03 | .76 | .09 | .874 |
| | Control vs. R3_Spaced | .10 | [.11, .63] | .03 | 2.88 | .37 | .021 |
| | R0 vs. R3_Mass | –.01 | [–.34, .25] | .04 | –.29 | .04 | .991 |
| | R0 vs. R3_Spaced | .06 | [–.07, .53] | .04 | 1.56 | .23 | .403 |
| | R3_Mass vs. R3_Spaced | .07 | [–.02, .57] | .04 | 1.87 | .27 | .244 |

*Note*: CI = confidence interval; Control = nonce-collocation controls.

practice rather than the retrieval practice alone. However, at the delayed posttest, the collocation processing advantage was observed only in the R3_Spaced condition ($d$ = .37 [small effect]). Additionally, the determiner presence had a significant effect ($d$ = .2 [small effect]) on the judgment times. The findings of the acceptability judgment posttest (i.e., the difference in RTs between the nonce collocations and target collocations) can be summarized as follows (note: PA = processing advantage):

Near-immediate posttest: R3_Spaced = R3_Massed = R0
Delayed posttest: R3_Spaced (PA) > R3_Massed (no PA) = R0

## *Results of primed lexical decisions*

### *Preliminary analysis*

In this analysis (see Table 6), we tested whether collocation priming (i.e., faster processing of the terminal word in the collocations than in the matched nonce controls) was observed for the familiar collocations. The familiar collocations showed no priming in the unstudied condition ($p$ = .648) or the studied condition ($p$ = .09). This suggests

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

**Table 6.** Response times of primed lexical decisions (familiar collocations): Fixed effects

| | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | –1.57 | [–1.75, –1.41] | .09 | –18.44 | –3.88 | < .001 |
| Exposure = Unstudied Familiar Collocation | –.02 | [–.08, .05] | .03 | –.46 | –.04 | .648 |
| Exposure = Studied Familiar Collocation | –.06 | [–.12, .01] | .03 | –1.7 | –.14 | .09 |
| Test_time = Delayed Posttest | –.27 | [–.47, –.07] | .1 | –2.73 | –.66 | .011 |

*Note*: CI = confidence interval; Reference level: Exposure = nonce-collocation control; Test_time = near-immediate posttest. Model formula: inverseRT ~ Exposure + Test_time + (Test_time + 1|Participant) + (1|Target).

that any gains in implicit knowledge of the target collocations in the primary analysis would likely reflect the effect of retrieval practice. The results showed a significant effect of Test_time ($d$ = .66 [medium effect]): the participants judged the stimuli (both the familiar collocations and their controls) faster at the delayed posttest than they did at the near-immediate posttest. However, this speed-up did not result in priming (as indicated by the absence of significant Exposure × Posttest interactions).

### *Primary analysis*

The Treatment × Test_time interaction was not a significant predictor of RT in this analysis and was not included in the final model (Table 7).

The post hoc comparisons of RTs for the levels of the learning condition and experimental condition (Table 8) showed that, after applying the Tukey adjustment, there was significant collocation priming in the R3_Spaced condition ($p$ < .05, $d$ = .23 [small effect]) but not in the R3_Massed condition ($p$ = .218, $d$ = .14 [small effect]), when collapsed across the posttests. As in the preliminary analysis, Test_time had a significant fixed effect ($d$ = .63 [medium effect]), with faster responses on the delayed

**Table 7.** Response times of primed lexical decisions (target collocations): Fixed effects

| | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | –1.52 | [–1.69, –1.35] | .09 | –17.48 | –4.03 | < .001 |
| Treatment = R0 | –.03 | [–.08, .03] | .03 | –.88 | –.07 | .377 |
| Treatment = R3_Mass | –.05 | [–.11, <.01] | .03 | –1.93 | –.14 | .054 |
| Treatment = R3_Spaced | –.09 | [–.14, –.03] | .03 | –3.06 | –.23 | .002 |
| Test_time = Delayed Posttest | –.24 | [–.46, –.02] | .11 | –2.15 | –.63 | .041 |
| Determiner = Yes | .06 | [–.01, .12] | .03 | 1.76 | .15 | .088 |

*Note*: CI = confidence interval; Reference level: Treatment = nonce-collocation control; Test_time = near-immediate posttest; Determiner = no determiner. Model formula: inverseRT ~ Treatment + Test_time + Determiner + (Test_time +1| Participant) + (1|Target).

**Table 8.** Post hoc comparisons of the response times for the levels of Treatment in the analysis of primed lexical decisions (target collocations)

| Contrast | *b* | 95% CI | *SE* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| Control vs. R0 | .03 | [–.09, .22] | .03 | .88 | .07 | .814 |
| Control vs. R3_Mass | .05 | [–.01, .3] | .03 | 1.93 | .14 | .218 |
| Control vs. R3_Spaced | .09 | [.08, .38] | .03 | 3.06 | .23 | .012 |
| R0 vs. R3_Mass | .03 | [–.11, .27] | .03 | 084 | .08 | .835 |
| R0 vs. R3_Spaced | .06 | [–.03, .35] | .03 | 1.75 | .16 | .301 |
| R3_Mass vs. R3_Spaced | .03 | [–.1, .27] | .03 | .91 | .08 | .801 |

*Note*: CI = confidence interval; Control = nonce-collocation controls.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

posttest. The findings of the primed lexical decision task (i.e., the difference in RTs between the nonce collocations and target collocations) can be summarized as follows:

R3_Spaced (significant priming) > R3_Massed (nonsignificant priming) > R0 (no priming)

## Discussion

The present study investigated the effect of retrieval schedules on the acquisition of implicit, automatized explicit, and explicit knowledge of L2 collocations. We found that retrieval schedules differentially affected the development of the three aspects of knowledge tested: (a) spaced and massed retrieval were equally beneficial for the acquisition of both the offline explicit knowledge (in the near-immediate and delayed posttests of form recall) and the initial automatized explicit knowledge (in the near-immediate posttest of acceptability judgment) and (b) spaced retrieval was more beneficial than massed retrieval for the retention of the automatized explicit knowledge (in the delayed posttest) and the acquisition of implicit collocation knowledge (in the near-immediate and delayed posttest of primed lexical decision). In other words, we observed the spacing effect in the retention of the automatized explicit knowledge and in the development of implicit collocational knowledge.

Surprisingly, the results of form recall showed that the massed and spaced retrieval were equally effective in promoting the acquisition of explicit knowledge, with both conditions yielding better results than the non-retrieval baseline condition. Numerically, spaced retrieval (near-immediate, 46%; delayed, 23%) led to slightly higher accuracy than massed retrieval (near-immediate, 44%; delayed, 20%), although there was no statistically significant difference between the two retrieval schedules. This finding is not aligned with the predicted spacing advantage in intentional MWE learning (Macis et al., 2021; Yamagata et al., 2023). The absence of the spacing effect in our study may be due to the measure of explicit knowledge, i.e., decontextualized *form* recall, which does not necessarily require access to meaning. It is possible that massed (consecutive) retrieval practice was particularly beneficial in creating a salient representation of the whole form of the collocations. The significant benefits of spaced and massed retrieval for explicit knowledge in our study may have also resulted from transfer-appropriate processing (Lightbown, 2007; Morris et al., 1977), as an overlap in the practice and test formats could have boosted response accuracy in the posttests. The transfer-appropriate processing effect, thus, may have offset the spacing effect (Van den Broek et al., 2018; Veltre et al., 2015).

An important new finding of our study concerns the acquisition of automatized explicit knowledge. We observed learning benefits for massed and spaced retrieval in the near-immediate posttest (i.e., initial learning). However, after 1 week, only spaced (but not massed) retrieval retained a processing advantage of the target collocations over the controls, indicating that the retention of the automatized explicit collocational knowledge exhibits the spacing effect. This finding also suggests that automatization of explicit knowledge of L2 collocations is not an all-or-nothing process, reflecting gradual performance improvement that may not be retained (Suzuki & DeKeyser, 2017). This finding is consistent with the previous finding that longer gaps between episodes tend to result in longer retention than shorter gaps. Still, shorter gaps may be either more beneficial (Cepeda et al., 2008) or equally effective (Kim & Webb, 2022) compared with longer gaps, when tested at short retention intervals. It is

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

also consistent with the prediction that more processing required in spaced retrieval (with feedback) is likely to benefit knowledge retention (Hintzman, 1976). Thus, any advantage of spaced retrieval may be more pronounced for long-term knowledge retention than initial learning, as predicted by the desirable difficulties account (Bjork, 1994; Schmidt & Bjork, 1992).

We also observed an advantage of spaced over massed retrieval for the development of implicit knowledge. A statistically significant collocation priming effect was recorded in the spaced retrieval condition, and the effect size of priming was larger in the spaced practice condition ($d$ = .25) than in the massed practice condition ($d$ = .16). Similar to Toomer and Elgort (2019), who reported a small effect size ($d$ = .15) for the collocation priming (in incidental learning), the priming effect sizes in our study were small. Suzuki and DeKeyser (2017) argue that the effects of spacing may be mediated by practice complexity and conditions that involve different cognitive processes. In our study, the retrieval practice was decontextualized, and only three retrieval opportunities were provided, which may have resulted in relatively weak implicit associations between the components of the target collocations. Perhaps a higher number of retrieval opportunities *in context* could lead to more robust implicit L2 collocational knowledge (Toomer & Elgort, 2019).

Our study shows that three instances of retrieval practice (after the familiarization stage) were sufficient for L2 learners to gain implicit knowledge of collocations (operationalized as collocation priming) in the spaced practice condition; in the massed practice condition, the implicit knowledge was still fragile (adjusted $p$ = .14). Sonbul and Schmitt (2013) did not find collocation priming in the decontextualized learning condition after three repetitions. This discord may be due to the differences in the two studies' learning procedures. In Sonbul and Schmitt's study, participants saw the target L2 collocations flashed on the screen in the decontextualized condition three times, for a *total* of 10 s, without meaning explanations or opportunities for retrieval. This learning treatment was probably insufficient for developing implicit collocational knowledge. In our study, the familiarization stage (where intact collocations were presented with their meanings and examples of use) likely resulted in the establishment of form-meaning mappings and whole phrase representations, strengthened by the subsequent form-focused retrieval practice and feedback; this practice procedure could have promoted the development of implicit associations between the component words of the target collocations. The finding that collocation priming was not statistically significant in the massed retrieval condition suggests that not all types of deliberate retrieval practice are equally effective for the development of implicit collocational knowledge. Our findings add to the limited research on the development of implicit knowledge of L2 collocations (Sonbul & Schmitt, 2013; Toomer & Elgort, 2019), suggesting that spaced retrieval combined with deliberate learning may benefit this knowledge type.

Finally, we discuss the findings for the familiar L2 collocations. We did not observe collocation priming for familiar collocations (either studied or unstudied). However, we did find a collocation advantage for both studied and unstudied familiar collocations in the acceptability judgment task. This suggests that (a) Chinese learners of English who participated in our study did not have observable implicit knowledge even of the L2 collocations that were considered known, but (b) their explicit knowledge of the familiar collocations was automatized. Our results show that, by engaging in spaced retrieval practice (with corrective feedback) after the familiarization stage, Chinese learners of English were able to develop not only automatized explicit knowledge but also implicit knowledge of L2 collocations.

https://doi.org/10.1017/S0272263124000184 Published online by Cambridge University Press

### Limitations

Several limitations to this study need to be acknowledged. First, the length of the first session might have introduced an element of fatigue among the participants. This could be mitigated in future studies by administering the tasks over more sessions and more days. We also acknowledge that the interval between the near-immediate and delayed posttests was relatively short (i.e., 1 week). Some previous studies used longer intervals, such as 2 weeks (e.g., Obermeier & Elgort, 2020; Sonbul & Schmitt, 2013), 3 weeks (e.g., Macis et al., 2021), and even 4 weeks (e.g., Farvardin, 2019). Longer intervals could be used in future studies to examine the effect of retrieval schedules on the longer-term retention of L2 collocational knowledge. Further, because the order in which the target collocations were presented in retrieval practice was kept the same, the delay between retrieval practice and near-immediate posttests was different for the target collocations; in future studies, it may be useful to randomize the order of the retrieval practice. Finally, the Vocabulary Levels Test (Webb et al., 2017) might provide a finer measure of the participants' lexical proficiency than LexTALE.

## Conclusions

The present study investigated the effects of retrieval schedules on the acquisition of explicit, automatized explicit, and implicit knowledge of L2 collocations. The results show that the effects of spacing vary by the type of collocational knowledge. Our findings corroborate the effectiveness of retrieval practice as a useful exercise for learning L2 MWEs (Strong & Boers, 2019a, 2019b). In addition, we found that spaced retrieval is more beneficial in learning L2 collocations than massed retrieval, adding to the existing evidence on the advantages of spaced retrieval practice on L2 single-word learning (e.g., Karatas et al., 2021) and further highlighting the relevance of the desirable difficulty framework of practice in the L2 field (Suzuki et al., 2019a, 2019b).

Our results show that the distinction between implicit knowledge and automatized explicit knowledge (Suzuki, 2017) is relevant in the study of L2 collocational knowledge. Our findings also show that retrieval schedules may differentially affect the development of offline and automatized explicit knowledge of L2 collocations. We, therefore, recommend instructional approaches that involve an initial presentation of intact collocations (such as the familiarization stage in our study) followed by multiple (ideally, more than three) *spaced* retrieval opportunities (with corrective feedback).

Importantly, the present study is only an initial step in researching the effect of retrieval schedules on the development of implicit and explicit knowledge of L2 collocations. Further research into the combined effects of retrieval schedules and frequency of retrieval episodes is theoretically and pedagogically interesting, because it can help us chart the time course of the acquisition of different aspects of L2 collocational knowledge and develop more robust instructional approaches.

**Supplementary material.** The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000184.

**Acknowledgments.** We would like to express our sincere gratitude to the editors of *Studies in Second Language Acquisition* and the anonymous reviewers for their valuable suggestions and constructive comments on earlier versions of this article. This work was partially funded by Shaoguan University scientific research funds (404-9900064604).

## References


Anthony, L. (2014). *AntWordProfiler*. Waseda University. https://www.laurenceanthony.net/software

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*, *21*, 19–31.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

BNC Consortium. *BNC XML edition.* (2007). British National Corpus. http://www.natcorp.ox.ac.uk/

Boers, F., Dang, T.C.T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*, 362–380.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, *18*, 54–74.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*, 860–864.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Craik, F.I.M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F.I.M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*, 268–294.

Farvardin, M. T. (2019). Effects of spacing techniques on EFL learners' recognition and production of lexical collocations. *Indonesian Journal of Applied Linguistics*, *9*, 395–403.

Garnier, M. M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*, 645–666.

Hamann, S. (1990). Level of processing effects in conceptually driven implicit tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 970–977.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. *10*, pp. 47–91). Academic Press.

Howarth, P. A. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*, 24–44.

Isbell, D. R., & Rogers, J. (2021). Measuring implicit and explicit learning and knowledge. In P. M. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 304–313). Routledge, Taylor & Franics Group.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The Tenten corpus family. In *The 7th International Corpus Linguistics Conference CL* (pp. 125–127).

Jeong, H., & DeKeyser, R. (2023). Development of automaticity in processing L2 collocations: The roles of L1 collocational knowledge and practice condition. *Studies in Second Language Acquisition*, *45*, 930–954. https://doi.org/10.1017/S0272263122000547

Jiang, N. (2012). *Conducting reaction time research in second language studies*. Routledge.

Karatas, N. B., Özemir, O., Lovelett, J. T., Demir, B., Erkol, K., Veríssimo, J., *et al.* (2021). Improving second language vocabulary learning and retention by leveraging memory enhancement techniques: A multi-domain pedagogical approach. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688211053525

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*, 704–719.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633.

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*, 269–319.

Koval, N. G. (2022). Testing the reminding account of the lag effect in L2 vocabulary learning. *Applied Psycholinguistics*, *43*, 1–40.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*, 694–716.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' *English.* Language Learning, *61*, 647–672.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *Emmeans: Estimated marginal means, aka least-squares means.* R package. https://CRAN.R-project. org/package= emmeans

Lightbown, P. M. (2007). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27–44). Multilingual Matters.

Macis, M., Sonbul, S., & Alharbi, R. (2021). The effect of spacing on incidental and deliberate learning of L2 collocations. *System*, *103*, 102649. https://doi.org/10.1016/j.system.2021.102649

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McIntosh, C. (Ed.) (2015). *Oxford collocations dictionary (English-Chinese edition).* Oxford University Press (China) / Foreign Language Teaching and Research Press.

McLaughlin, B. (1987). *Theories of second-language learning*. Routledge.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*, 233–260.

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*, 287–311.

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*, 223–242.

Newell, B. R., & Andrews, S. (2004). Levels of processing effects on implicit and explicit memory tasks: Using question position to investigate the lexical-processing hypothesis. *Experimental Psychology*, *51*, 132–144.

Northbrook, J., Allen, D., & Conklin, K. (2022). 'Did you see that?'—The role of repetition and enhancement on lexical bundle processing in English learning materials. *Applied Linguistics*, *43*, 453–472.

Obermeier, A., & Elgort, I. (2020). Deliberate and contextual learning of L2 idioms: The effect of learning conditions on online processing. *System*, 102428.

Öksüz, D., Brezina, V., & Rebuschat, P. (2021). Collocational processing in L1 and L2: The effects of word frequency, collocational frequency, and association. *Language Learning*, *71*, 55–98.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 23–47). Psychology Press.

Roediger, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 3–41). Lawrence Erlbaum Associates, Inc.

Rogers, J. (2023). Spacing effects in task repetition research. *Language Learning*, *73*(2), 445–474.

Siyanova-Chanturia, A., & Van Lancker Sidtis, D. (2019). What online processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 38–61). Routledge.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–218.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*, 281–317.

Schmitt, N. (2022). Norbert Schmitt's essential bookshelf: Formulaic language. *Language Teaching*, *56*, 420–431. https://doi.org/10.1017/S0261444822000039

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*, 176–199.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*, 121–159.

Stengers, H., & Boers, F. (2015). Exercises on collocations: A comparison of trial-and-error and exemplar-guided procedures. *Journal of Spanish Language Teaching*, *2*, 152–164.

Strong, B., & Boers, F. (2019a). The error in trial and error: Exercises on phrasal verbs. *TESOL Quarterly*, *53*, 289–319.

Strong, B., & Boers, F. (2019b). Weighing up exercises on phrasal verbs: Retrieval versus trial-and-error practices. *The Modern Language Journal*, *103*, 562–579.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*, 166–188.

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019a). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *The Modern Language Journal*, *103*, 551–561.

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019b). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*, 713–720.

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics*, *26*, 245–265.

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*, 783–783.

Tsai, M.-H. (2020). The effects of explicit instruction on L2 learners' acquisition of verb–noun collocations. *Language Teaching Research*, *24*, 138–162.

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *34*, 39–65.

Van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, *68*, 546–585.

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*, 1229–1237.

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *TESOL Quarterly*, *43*, 55–77.

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics*, *32*, 259–276.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*, 33–69.

Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, *40*, 395–416.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Yamagata, S., Nakata, T., & Rogers, J. (2023). Effects of distributed practice on the acquisition of verb-noun collocations. *Studies in Second Language Acquisition*, *45*, 291–317. https://doi.org/10.1017/S0272263122000225

Yamashita, J., & Jiang, N. (2010). L1 Influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*, 647–668.

**Cite this article:** Fang, N., Elgort, I., & Chen, Z. (2024). Effects of retrieval schedules on the acquisition of explicit, automatized-explicit, and implicit knowledge of L2 collocations. *Studies in Second Language Acquisition,* 46: 663–685. https://doi.org/10.1017/S0272263124000184