Meaning recall was also more accurate in the immediate than the delayed posttest, and for the cooking than the building theme (Table 2).

Table 2. Accuracy of meaning recall (fixed effects)

| Parameter | *b* | *SE* | *z* | *p* | *d* |
|---|---|---|---|---|---|
| Intercept[a] | –4.09 | 0.32 | –12.94 | < .001 | –2.26 |
| Schedule=spaced | 1.55 | 0.27 | 5.81 | < .001 | 0.86 |
| Inference accuracy=1 | 0.10 | 0.23 | 0.45 | .652 | 0.06 |
| Session=immediate | 0.72 | 0.08 | 9.31 | < .001 | 0.40 |
| Theme=cooking | 0.67 | 0.28 | 2.37 | .018 | 0.37 |
| Schedule=spaced:Infer.accuracy=1 | 1.20 | 0.24 | 4.91 | < .001 | 0.66 |

*Note*. [a]Intercept levels: Schedule = massed, Inference accuracy = 0 (incorrect), Session = delayed, Theme = building.

**Analysis of Responses in the Meaning-form Matching Posttest**

In the analysis of meaning-form matching there was a significant interaction between Schedule and Inference accuracy (Table 3, Figure 3): in the spaced treatment, accuracy of meaning-form matching was better when the final inference was correct; inference accuracy did not affect accuracy of meaning-form matching in the massed treatment.

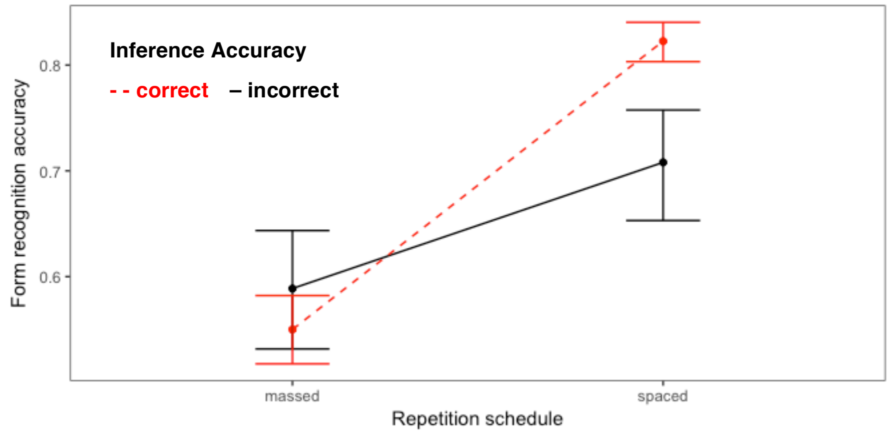


Figure 3. Effect plot for the interaction between Schedule and Inference accuracy in the analysis of meaning-form matching (fit and 95% CIs).

Importantly, there was also a main effect of Schedule ($z$ = 2.79, $p$ = .005), with around 24% advantage for the spaced over massed treatment (based on the model fit). Meaning recall was more accurate in the immediate than the delayed posttest, and for the cooking than building theme (Table 3).

Table 3. Analysis of the accuracy of meaning-form matching (fixed effects)

| Parameter | *b* | *SE* | *z* | *p* | *d* |
|---|---|---|---|---|---|
| Intercept[a] | 0.57 | 0.21 | 2.73 | .006 | |
| Schedule=spaced | 0.53 | 0.19 | 2.79 | .005 | 0.29 |
| Inference accuracy=1 | –0.16 | 0.15 | –1.07 | .285 | –0.09 |
| Theme=home | –0.42 | 0.20 | –2.13 | .034 | –0.23 |
| Schedule=spaced: Infer.accuracy =1 | 0.81 | 0.21 | 3.83 | < .001 | 0.45 |

*Note*. [a]Intercept levels: Schedule = massed, Inference accuracy = 0 (incorrect), Session = delayed, Theme = building.

**Analysis of Responses in the Semantic Priming Task**

We first conducted the analyses of the accuracy and response times data of lexical decisions on the word targets. In these analyses, the experimental condition (i.e., related/unrelated) was treated as a primary interest predictor. We predicted faster responses to the word targets in the related condition compared to the unrelated condition (i.e., priming) for the pseudowords that had been integrated into the lexical semantic networks of the learners. The second primary predictor was Schedule. If the spacing schedule differentially affected the development of semantic knowledge of the pseudowords, we would expect to see an interaction between the experiential condition (Cond) and Schedule. In addition to the analysis of the lexical decisions to the target words, we also analyzed lexical decisions to the primes (i.e., the pseudowords and real words).

***Semantic priming: Accuracy analysis***

The analysis showed significant negative semantic priming ($z = 2.79$, $p = .003$), with responses in the unrelated condition being about 2% more accurate than on semantically related trials, but there was no significant interaction between priming and Schedule (Appendix B). There was no effect of inference accuracy in this analysis.

***Semantic priming: Response time analysis***

Incorrect responses to targets (and their corresponding primes) were removed prior to the data analysis (21% of the data points). The final model included three significant two-way interactions: (a) between the experimental condition and accuracy of lexical decisions to the prime, (b) between the experimental condition and speed of lexical decisions to the prime, and (c) between accuracy of lexical decisions to the prime and the posttest session (immediate/delayed) (Table 4). Importantly, there was also a main effect of the experimental condition: responses in the related condition were faster than in the unrelated condition; that is, we observed the semantic priming effect for the pseudowords. However, there was no effect of Schedule, nor was there an interaction between Schedule and the experimental condition.

Table 4. Semantic priming, response time analysis (fixed effects)

| Parameter | *b* | *SE* | *df* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| Intercept[a] | –1.25 | 0.04 | 105.52 | –34.57 | < .001 | 3.25 |
| Cond=unrelated | 0.16 | 0.02 | 994.69 | 6.81 | < .001 | 0.41 |
| Prime RT (inv.Prime.RT.c[b]) | 0.07 | 0.02 | 145.83 | 3.76 | < .001 | 0.18 |
| Prime accuracy (Prime.ACC=1) | –0.02 | 0.01 | 4403.99 | –1.44 | .151 | 0.06 |
| Session=immediate | 0.06 | 0.02 | 135.49 | 3.96 | < .001 | 0.17 |
| Target length (centered) | 0.08 | 0.01 | 46.96 | 5.67 | < .001 | 0.20 |
| Cond= unrelated:invPrime.RT.c | 0.07 | 0.02 | 1037.73 | 3.27 | .001 | 0.18 |
| Cond=unrelated:Prime.ACC=1 | –0.11 | 0.02 | 4596.23 | –4.55 | < .001 | 0.29 |
| Prime.ACC=1:Session=immediate | –0.06 | 0.02 | 4674.36 | –4.21 | < .001 | 0.17 |

*Note*. [a]Intercept levels: Condition = related, Prime accuracy = 0, Session = delayed.

[b]Response times to primes, inverse transformed and centered.

***Lexical decisions to primes: Response accuracy analysis***

There was a significant two-way interaction in the analysis of response accuracy between Item type and Schedule; as expected, responses to the known words were not affected by the spacing schedule, but responses to the pseudowords were (Figure 4, Table 5). The lexical decisions were more accurate when the pseudowords had been encountered in the spaced than in the massed treatment.

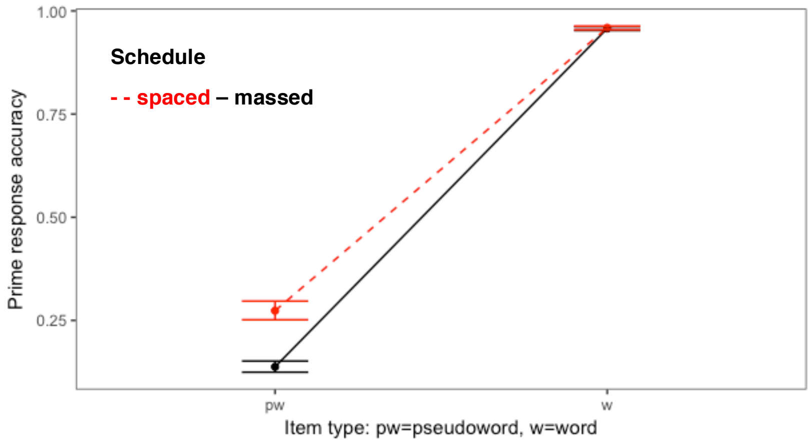


Figure 4. Effect plot for the interaction between Schedule and Item type in the analysis of accuracy of lexical decisions to primes (fit and 95% CIs).

Table 5. Lexical decisions to primes, response accuracy (fixed effects)

| Parameter | *b* | *SE* | *z* | *p* | *d* |
|---|---|---|---|---|---|
| Intercept[a] | –2.04 | 0.19 | –11.00 | < .001 | –1.13 |
| Item type=word | 4.95 | 0.23 | 21.08 | < .001 | 2.73 |
| Schedule=spaced | 0.86 | 0.09 | 9.19 | < .001 | 0.47 |
| Session=immediate | 0.41 | 0.08 | 5.32 | < .001 | 0.23 |
| Item type=word:Schedule=spaced | –0.82 | 0.17 | –4.79 | < .001 | –0.45 |

*Note*. [a]Intercept levels: Item type = pseudoword, Schedule = massed, Session = delayed.



***Lexical decisions to primes: Response time analysis***

Incorrect responses to primes were not removed from the analysis; instead, prime accuracy was included in the model. There was no effect of Schedule in the analysis of response times to primes (for the results, see Appendix C). The only result of note in this analysis was a two-way interaction between Item type and Session ($t = –8.48$, $p < .001$, $d = 0.25$); response times to the word primes were similar in the immediate and delayed sessions but responses to the pseudowords were faster in the delayed session.

**Summary of Findings**

We have compared the effect of two repetition schedules on contextual word learning. In the massed schedule, participants first read all three sentences with the target pseudoword presented on the same screen (no lag), then inferred its meaning and then reviewed the correct meaning. In the spaced condition, the participants read each of the three sentences separately with a lag of about 25 minutes; they inferred the meaning of the pseudoword after each encounter and reviewed the correct meaning after each inference attempt. We found that the participants were better able to infer the meanings of the pseudowords at the end of the spaced treatment than in the massed treatment. However, the first and the second meaning inference attempts in the spaced treatment were less accurate than in the massed treatment. This suggests that spaced contextual word learning is incremental and is facilitated through an inference-feedback loop that involves making contextual inferences, retrieving knowledge from previous learning episodes, and processing feedback.

Both explicit knowledge posttests (meaning-form matching and meaning recall) exhibited the spacing effect: better learning and retention was observed in the spaced than in the massed treatment, especially when the last inference in the spaced treatment was correct. Lexical decisions to the pseudoword primes were also more accurate when they were encountered in the spaced treatment, pointing to more precise lexical representations. We did not observe the spacing effect on the tacit knowledge, operationalized as semantic priming.

## Discussion

In the present study, spaced distribution led to significantly higher scores than massed distribution on explicit knowledge posttests (meaning-form matching and meaning recall). The findings suggest that the spacing effect can be observed not only in decontextualized but also contextual vocabulary learning. The advantage of the spaced distribution over massed distribution might also be due in part to increased opportunities for retrieval and incremental access to feedback after each contextual inference. In the spaced condition, three sentences involving a given pseudoword were presented one by one, after approximately 25 minutes each. Each contextual encounter was accompanied by an explicit inference attempt and feedback. The second and third presentations also allowed learners to retrieve information about the pseudoword learned in previous encounters. In the massed condition, in contrast, three sentences for a given pseudoword were presented simultaneously and constituted only one learning episode with one explicit inference attempt followed by feedback. Thus, the massed condition provided considerably reduced retrieval opportunities rather than a full inference-retrieval-feedback learning cycle. The superiority of spaced distribution over massed distribution in this study is consistent with earlier research demonstrating positive effects of retrieval and feedback on L2 vocabulary learning (e.g., Barcroft, 2007; Karpicke and Roediger, 2008; van den Broek et al., 2018).

Although spaced distribution was significantly more effective than massed distribution on posttests measuring explicit knowledge (meaning-form matching and meaning recall), the present study demonstrated no significant advantage of spaced distribution for the acquisition of tacit knowledge. The differential effect of the repetition schedule on type of knowledge in this study confirms that the acquisition of explicit and tacit knowledge may be affected by different factors (Bird, 2010; Elgort et al., 2018; Elgort and Warren, 2014; Suzuki and DeKeyser, 2017).

Another explanation for the lack of spacing effect on the acquisition of tacit knowledge is that the simultaneous presentation of three sentences in the massed condition may have encouraged (re)activation of the core senses of the pseudoword presented in three different informative contexts on the same screen, strengthening semantic connections of the novel vocabulary item with known L2 words, thus offsetting the spacing effect. In other words, although the spaced condition resulted in larger gains in the knowledge of explicit form-meaning mapping (measured by the posttests that afford explicit retrieval approaches), no spacing effect was observed in the semantic priming task relying on the online activation of semantic connections. This result is in line with the instance-based framework of word learning (Bolger et al., 2008), which suggests that multiple encounters with a word in supportive contexts result in the establishment of its semantic representation, as the word's core semantic features become abstracted from specific contexts.

Notably, the advantage of spacing was not observed in initial stages of learning. The first and second meaning inference attempts in the spaced treatment were less accurate than in the massed treatment, with only the third, final meaning inference being more accurate in the spaced than massed treatment. This corroborates the finding from the contextual word learning literature that multiple encounters are needed to acquire a word from reading (Elgort et al., 2018; Elgort and Warren, 2014; Pellicer-Sánchez, 2016; Pellicer-Sánchez and Schmitt, 2010; Webb, 2007). Our study shows contextual word learning both as a dynamic process and an outcome, for the two repetition schedules. The initial superior accuracy of meaning inferences in the massed condition may have been due to a more successful inductive learning from multiple simultaneously-presented examples of item use in context, providing more contextual clues than a single encounter with the word. This advantage diminished, however, by the second occurrence of the item in the spaced schedule, and was reversed by the third, as a result of the repeated distributed inference and retrieval attempts followed by feedback. At posttest, the advantage of the spaced treatment for the development of explicit knowledge of meaning and form-meaning mapping was clear.

We also observed a significant interaction between inference accuracy and schedule on the meaning recall and meaning-form matching posttests. In the spaced condition, the correct response on the last (third) inference attempt during the learning phase was associated with the correct response on the posttest, whereas there was no significant relationship between the inference accuracy and posttest performance in the massed condition. One explanation is the differential effects of retrieval and inference success on vocabulary learning. Memory research suggests that successful retrieval leads to better retention than unsuccessful retrieval because successfully recalling information strengthens a retrieval route, facilitating subsequent retrieval (e.g., Baddeley, 1997). However, the provision of feedback counteracts possible negative effects of initial incorrect retrieval (Carpenter et al., 2012; Elgort et al., 2020). Recall that the correct response on the inference attempt in the massed condition was due purely to inference success whereas the correct response on the third attempt in the spaced condition may have been caused by a combination of inference and retrieval success and the effect of feedback. This may explain why the learners' posttest performance in the spaced, but not massed, condition was mediated by the accuracy of the final inference in the learning phase.

Another plausible explanation is that the inference accuracy for the third retrieval attempt in the spaced condition reflects the learning burden of the pseudowords whereas the single inference attempt in the massed condition does not. Because in the spaced condition the learners reviewed the correct meaning of the pseudoword after submitting a contextual meaning inference twice before making the final, third inference, an inference error on the third attempt suggests that the pseudoword was perhaps more difficult to learn, i.e., was associated with a greater learning burden. In the massed condition, learners were not exposed to the correct meanings of the pseudowords prior to the single inference attempt; therefore, inference accuracy in the massed treatment indexed the guessability (ease of meaning inferences) rather than the learnability of the pseudowords. This may explain why the accuracy of explicit pseudoword knowledge on the posttests was associated with the inference accuracy in the spaced but not massed learning condition.

**Theoretical Account of the Findings**

A number of theoretical frameworks have been proposed to explain the spacing effect, such as the encoding variability account (e.g., Maddox, 2016), deficient processing account (e.g., Koval, 2019), and transfer appropriate processing account (e.g., Russo & Mammarella, 2002). According to the encoding variability account, in a massed schedule, information tends to be encoded in a stable, fixed context, whereas in a spaced schedule, it is encoded in more physically, mentally, or temporally diverse contexts. The encoding variability account suggests that more diverse encoding processes in a spaced schedule facilitate later recall. The deficient processing account states that information presented in a spaced schedule receives more attention or processing than information presented in a massed schedule, which results in superior retention. Transfer appropriate processing theory suggests that memory performance is enhanced if there is a close match between the context of encoding and that of testing (Morris, Bransford, & Franks, 1977). When applied to the spacing effect, this theory predicts that spaced learning, where information is presented over a longer period of time, results in the kind of knowledge that can be accessible for a long time, whereas massed learning may facilitate only short-term memory.

The findings of the present study cannot be fully accounted for by either the encoding variability or deficient processing account because neither of the two frameworks predicts the advantage of spacing over massing for the acquisition of explicit, but not tacit knowledge. The results of this study may be better explained by the transfer appropriate processing account. In the spaced (but not in the massed) condition, participants had multiple, distributed opportunities to retrieve the meaning of the previously encountered target items. Note that explicit measures used in this study (i.e., meaning recall and meaning-form matching posttests) also required learners to retrieve the meaning of the target items. However, neither spaced nor massed sentence reading treatment required the kind of processing tapped into by the semantic priming task, that is, implicit processing of the target items and words related to them in their meaning (rather than words that co-occurred with them in the text). The transfer appropriate processing account, therefore, predicts better performance for the spaced than massed condition on the posttests of explicit knowledge of meaning, but not on the measure of tacit knowledge of meaning, as is the case in the present study. However, because the transfer appropriate account of the spacing effect has not received as much attention from researchers as the encoding variability or deficient processing account, more empirical studies need to be conducted to test the validity and explanatory power of this account.

**Practical Implications of the Findings**

A key finding of our study is that, in L2 vocabulary acquisition from reading, the development of explicit knowledge of form and meaning (and their mapping) is facilitated when new words occur with a lag rather than when they are clumped together within the text, particularly when learners are able to verify their contextual inferences, e.g., using a dictionary or glossary. This insight is useful for content developers (both publishers and teachers) because it can help them create or select more effective materials for L2 learners to build their vocabulary from reading. In addition, we found that tacit semantic knowledge develops from encountering new words in supportive context, irrespective of whether the new words occur in a massed or spaced manner in the text, which emphasizes the importance of reading for vocabulary development.

Although some researchers argue that pure massing rarely takes place in a classroom setting and is not very educationally relevant (e.g., Rogers and Cheung, in press), we maintain that massing does in fact occur in L2 teaching. For example, in concordance-based learning activities recommended by a number of L2 vocabulary scholars (e.g., Nation, 2013; Schmitt, 2000), multiple instances of a new word are presented in context on the same page. Thus, learning words from a concordance output is essentially massed learning. Although concordance-based activities can help learners notice patterns of language use (for example, how words co-occur) or generate explicit meaning inferences, our findings suggest that concordance outputs will be less effective in learning form-meaning connections. Spaced encounters with new words in supportive contexts are likely to result in better learning outcomes, at least for the learning of form-meaning connections.

Massing is also a common way of presenting items in contextual L2 vocabulary learning research. Based on the findings of the present study, this may not be an optimal learning choice. We recommend that future vocabulary learning studies deliberately consider the effect of spacing at the research design stage.

**Concluding Remarks**

This study examined the effects of massed and spaced distributions on the acquisition of explicit and tacit vocabulary knowledge from reading. Our study design emulated an authentic learning scenario in which L2 learners verify their contextual meaning inferences of unknown words by consulting dictionaries, asking their instructors or peers (Schmitt, 1997), or using glossaries. The results showed that, when contextual meaning inferencing is required and is followed by feedback (i.e., L2 synonyms and L1 translation equivalents), spaced learning is more effective for the acquisition of explicit knowledge of form-meaning mapping than massed learning. However, the spacing advantage may not hold without the provision of feedback after each inference attempt. A combination of spaced retrieval and feedback is known to facilitate L2 vocabulary learning (e.g., Barcroft, 2007; Karpicke and Roediger, 2008; van den Broek et al., 2018); therefore, our design may have worked in favor of the spaced treatment, which involved opportunities for retrieval. In future research, the effects of massing and spacing on contextual vocabulary learning should also be examined without access to feedback in order to verify whether vocabulary learning from reading without access to reference materials is differentially affected by spacing.

Another useful follow-up from this study would be an investigation of how spacing affects tacit knowledge in decontextualized paired-associate learning. In our study, spacing did not affect the development of tacit semantic knowledge. We hypothesized that, by presenting multiple sentences simultaneously, massing may have allowed learners to establish a more robust set of semantic features that they could access in the online semantic priming posttest. In decontextualized learning, however, massing does not offer the same advantage; thus spacing should be more effective than massing for the acquisition of tacit semantic knowledge. Considering that existing decontextualized studies of the spacing effect only measured explicit knowledge (e.g., Karpicke and Bauernschmidt, 2011; Nakata, 2015; Seibert, 1932), further research comparing the effects of massing and spacing on the acquisition of tacit semantic knowledge in decontextualized learning is warranted.

Compared with the lag effect, which is sometimes referred to as "nebulous" (Rogers, 2017, p. 907), the spacing effect is considered very robust in associative learning, and L2 vocabulary researchers have recommended a spaced learning approach (e.g., Barcroft, 2015b; Ellis, 1995; Hulstijn, 2001; Nation, 2013; Schmitt, 2000).[3] Our study makes initial steps in extending the spacing effect to contextual vocabulary learning, as far as the acquisition of explicit knowledge is concerned. On the other hand, massing appears to be just as effective as spacing for the acquisition of tacit knowledge. Methodologically, our findings suggest that when comparing the effects of massing and spacing, it is important to measure both explicit and tacit knowledge because tacit knowledge of meaning underpins fluent word-to-text integration. Further research examining the effects of spacing on the acquisition of tacit knowledge is warranted because rich and flexible tacit semantic knowledge is essential for fluent language processing.

## Notes

[1] In some studies, the term *massed learning* is used to refer to relatively short spacing, whereas *spaced learning* refers to relatively long spacing. In this study, massed learning is used to refer to a practice schedule that does not involve any spacing (Cepeda et al., 2006).

[2] Hulstijn (2001) defines incidental vocabulary learning as an activity where participants are not explicitly instructed to learn target vocabulary items and are not aware of upcoming vocabulary posttests. Barcroft (2015b), however, argues that the lack of forewarning of vocabulary posttests does not necessarily guarantee that learners do not attempt to learn target vocabulary items intentionally and considers it more appropriate to use the term *incidentally-oriented vocabulary learning* instead of incidental vocabulary learning.

[3] For non-L2 vocabulary research failing to show the advantage of spacing over massing (i.e., *Peterson paradox*), see a meta-analysis conducted by Cepeda and colleagues (2006).

**Acknowledgements**

An earlier version of this paper was presented at Vocab@Leuven Conference and Japan Second Language Acquisition Research Forum Second Meeting. This research was supported in part by JSPS Grant-in-Aid for Research (#16H05943) awarded to the first author. We are very grateful to Shotaro Ueno, Saki Nagamine, and Hiroko Kashiba for their assistance with data collection. We are also very grateful to two anonymous reviewers, Anna Siyanova, and the SLR Editors for their invaluable feedback on earlier versions of the manuscript.



## References

Baayen RH (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK, Cambridge University Press.

Baayen RH, Davidson DJ and Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Baddeley AD (1997) *Human memory: Theory and practice*. Revised ed. East Sussex, UK, Psychology Press.

Bahrick HP and Phelps E (1987) Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 344–349.

Barcroft J (2007) Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57, 35–56.

Barcroft J (2015a) Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, 48, 236–249.

Barcroft J (2015b) *Lexical input processing and vocabulary learning*. Amsterdam, Netherlands, John Benjamins.

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

Bird S (2010) Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, 31, 635–650.

Bolger DJ, Balass M, Landen E, et al. (2008) Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45, 122–159.

Brysbaert M and Stevens M (2018) Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1, 9.

Carpenter SK, Sachs RE, Martin B, et al. (2012) Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin and Review*, 19, 81–86.

Cepeda NJ, Pashler H, Vul E, et al. (2006) Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354–380.

Cunnings I (2012) An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28, 369–382.

Ebbinghaus H (1885) Memory: A contribution to experimental psychology. New York, NY, Genesis Publishing.

Elgort I (2011) Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61, 367–413.

Elgort I (2017) Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, 1, 1–11.

Elgort I, Brysbaert M, Stevens M, et al. (2018) Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40, 341–366.

Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition, 42*, 7-32. doi:10.1017/S0272263119000561

Elgort I and Warren P (2014) L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, 64, 365–414.

Ellis NC (1995) The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, 8, 103–128.

Hothorn T, Bretz F and Westfall P (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, 50, 346–363.

Hulstijn JH (2001) Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In: Robinson P (ed.), *Cognition and second language instruction*, Cambridge, UK, Cambridge University Press, pp. 258–286.

Karpicke JD and Bauernschmidt A (2011) Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1250–1257.

Karpicke JD and Roediger HL (2008) The critical importance of retrieval for learning. *Science*, 319, 966–968.

Kobayashi A and Okubo M (2014) 日本語版オペレーションスパンテストによるワーキングメモリの測定(Assessment of working memory capacity with a Japanese version of the Operation Span Test). *The Japanese Journal of Psychology*, 85, 60–68.

Kornell N and Bjork RA (2008) Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, 19, 585–592.

Koval NG (2019) Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, 40, 1103–1139.

Kurtz KH and Hovland CI (1956) Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, 51, 239–243.

Laufer B (2003) Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, 59, 567–587.

Laufer B and Hulstijn JH (2001) Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22, 1–26.

Maddox GB (2016) Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, 28, 684–706.

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

McRae K and Boisvert S (1998) Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 558–572.

Metcalfe J and Kornell N (2007) Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin and Review*, 14, 225–229.

Morris CD, Bransford JD and Franks JJ (1977) Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16, 519–533.

Nakata T (2015) Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37, 677–711.

Nakata T and 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 ISP (2013) *Learning vocabulary in another language*. 2nd ed. Cambridge, UK, Cambridge University Press.

Nation ISP and Beglar D (2007) A vocabulary size test. *The Language Teacher*, 31(7), 9–13.

Pellicer-Sánchez A (2016) Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, 38, 97–130.

Pellicer-Sánchez A and Schmitt N (2010) Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, 22, 31–55.

Rogers J (2017) The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, 38, 906–911.

Rogers J and Cheung A (in press) Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*.

Russo R and Mammarella N (2002) Spacing effects in recognition memory: When meaning matters. *European Journal of Cognitive Psychology*, 14, 49–59.

Schmitt N (1997) Vocabulary learning strategies. In: Schmitt N and McCarthy M (eds.), *Vocabulary: Description, acquisition and pedagogy*, Cambridge, UK, Cambridge University Press, pp. 199–227.

Schmitt N (2000) *Vocabulary in language teaching.* Cambridge, UK, Cambridge University Press.

Seibert LC (1932) *A series of experiments on the learning of French vocabulary*. Baltimore, MD, The Johns Hopkins Press.

Serrano R and Huang H-Y (2018) Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, 52, 971–994.

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

Tinkham T (1997) The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research,* 13, 138–163.

van den Broek GSE, Takashima A, Segers E, et al. (2018) Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, 68, 546–585.

Webb S (2007) The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28, 46–65.

Webb S and Chang A (2015) Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, 19, 667–686.

Yanagisawa A, Webb S and Uchihara T (in press) How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*.

Zulkiply N and Burt JS (2013) The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory and Cognition*, 41, 16–27.

**Appendix A. Sample item order during the treatment**

The following figure gives a sample item order during the treatment. In the figure, "Massed 1" refers to the first pseudoword assigned to the massed condition, and "Spaced 1" refers to the first pseudoword assigned to the spaced condition.

| Position | Item | Sentences |
|---|---|---|
| 1 | Massed 1 | 1, 2, 3 |
| 2 | Spaced 1 | 1 |
| 3 | Spaced 2 | 1 |
| 4 | Spaced 3 | 1 |
| 5 | Massed 2 | 1, 2, 3 |
| 6 | Spaced 4 | 1 |
| 7 | Spaced 5 | 1 |
| 8 | Spaced 6 | 1 |
| 9 | Massed 3 | 1, 2, 3 |
| 10 | Spaced 7 | 1 |
| 11 | Spaced 8 | 1 |
| 12 | Spaced 9 | 1 |
| 13 | Massed 4 | 1, 2, 3 |
| 14 | Spaced 10 | 1 |
| 15 | Spaced 11 | 1 |
| 16 | Spaced 12 | 1 |
| 17 | Massed 5 | 1, 2, 3 |
| 18 | Spaced 13 | 1 |
| 19 | Spaced 14 | 1 |
| 20 | Spaced 15 | 1 |
| 21 | Massed 6 | 1, 2, 3 |
| 22 | Spaced 16 | 1 |
| 23 | Spaced 17 | 1 |
| 24 | Spaced 18 | 1 |
| 25 | Massed 7 | 1, 2, 3 |
| 26 | Spaced 19 | 1 |
| 27 | Spaced 20 | 1 |
| 28 | Spaced 21 | 1 |
| 29 | Massed 8 | 1, 2, 3 |
| 30 | Spaced 22 | 1 |
| 31 | Spaced 23 | 1 |
| 32 | Spaced 24 | 1 |
| 33 | Massed 9 | 1, 2, 3 |
| 34 | Spaced 1 | 2 |
| 35 | Spaced 2 | 2 |
| 36 | Spaced 3 | 2 |
| 37 | Massed 10 | 1, 2, 3 |
| 38 | Spaced 4 | 2 |
| 39 | Spaced 5 | 2 |
| 40 | Spaced 6 | 2 |
| 41 | Massed 11 | 1, 2, 3 |
| 42 | Spaced 7 | 2 |
| 43 | Spaced 8 | 2 |
| 44 | Spaced 9 | 2 |
| 45 | Massed 12 | 1, 2, 3 |
| 46 | Spaced 10 | 2 |
| 47 | Spaced 11 | 2 |
| 48 | Spaced 12 | 2 |
| 49 | Massed 13 | 1, 2, 3 |
| 50 | Spaced 13 | 2 |
| 51 | Spaced 14 | 2 |
| 52 | Spaced 15 | 2 |
| 53 | Massed 14 | 1, 2, 3 |
| 54 | Spaced 16 | 2 |
| 55 | Spaced 17 | 2 |
| 56 | Spaced 18 | 2 |
| 57 | Massed 15 | 1, 2, 3 |
| 58 | Spaced 19 | 2 |
| 59 | Spaced 20 | 2 |
| 60 | Spaced 21 | 2 |
| 61 | Massed 16 | 1, 2, 3 |
| 62 | Spaced 22 | 2 |
| 63 | Spaced 23 | 2 |
| 64 | Spaced 24 | 2 |
| 65 | Massed 17 | 1, 2, 3 |
| 66 | Spaced 1 | 3 |
| 67 | Spaced 2 | 3 |
| 68 | Spaced 3 | 3 |
| 69 | Massed 18 | 1, 2, 3 |
| 70 | Spaced 4 | 3 |
| 71 | Spaced 5 | 3 |

| | | |
|---|---|---|
| 72 | Spaced 6 | 3 |
| 73 | Massed 19 | 1, 2, 3 |
| 74 | Spaced 7 | 3 |
| 75 | Spaced 8 | 3 |
| 76 | Spaced 9 | 3 |
| 77 | Massed 20 | 1, 2, 3 |
| 78 | Spaced 10 | 3 |
| 79 | Spaced 11 | 3 |
| 80 | Spaced 12 | 3 |
| 81 | Massed 21 | 1, 2, 3 |
| 82 | Spaced 13 | 3 |
| 83 | Spaced 14 | 3 |
| 84 | Spaced 15 | 3 |
| 85 | Massed 22 | 1, 2, 3 |
| 86 | Spaced 16 | 3 |
| 87 | Spaced 17 | 3 |
| 88 | Spaced 18 | 3 |
| 89 | Massed 23 | 1, 2, 3 |
| 90 | Spaced 19 | 3 |
| 91 | Spaced 20 | 3 |
| 92 | Spaced 21 | 3 |
| 93 | Massed 24 | 1, 2, 3 |
| 94 | Spaced 22 | 3 |
| 95 | Spaced 23 | 3 |
| 96 | Spaced 24 | 3 |

As shown above, the three sentences from the massed and spaced conditions alternated throughout the treatment. For instance, at the beginning of the treatment, three sentences for the first item in the massed condition (Massed 1) were presented for 90 seconds. After that, three sentences from the first three items in the spaced condition (Spaced 1–3) were presented one by one, for 30 seconds each. This was followed by three sentences for the second item in the massed condition (Massed 2) presented for 90 seconds.

The item order was randomized anew for each participant to minimize the order effect. For instance, for one participant, *shottle* (gravel 砂利) may be Massed 1 and *tenont* (pumpkin かぼちゃ) may be Massed 24, whereas for another participant, *dapson* (detergent 洗剤) may be Massed 1 and *brophy* (strainer ざる) may be Massed 24. The randomization was implemented in such a way that pseudowords from the two themes (building/household and cooking/food) were distributed roughly equally across the treatment. To control for the order effect, for half of the participants, the first three sentences were from the massed condition, and for the other half of the participants, the first three sentences were from the spaced condition.

**Appendix B. Semantic priming, accuracy analysis (fixed effects)**

| Parameter | *b* | *SE* | *z* | *p* | *d* |
|---|---|---|---|---|---|
| Intercept[a] | 1.73 | 0.26 | 6.73 | < .001 | 0.95 |
| Cond=unrelated | 0.22 | 0.08 | 2.97 | .003 | 0.12 |
| Schedule=spaced | 0.20 | 0.08 | 2.63 | .009 | 0.11 |
| Session=immediate | 0.34 | 0.08 | 4.57 | < .001 | 0.19 |
| Vocabulary size (VST.lg.c[b]) | 4.01 | 1.06 | 3.77 | < .001 | 2.21 |
| Target length (Target.length.c[c]) | –0.30 | 0.12 | –2.64 | .008 | –0.17 |

*Note*. [a]Intercept levels: Condition = related, Schedule = massed, Session = delayed.

[b]Vocabulary size score, log-transformed, centered. [c]Target length in letters, centered.

**Appendix C. Lexical decisions to primes, response times (fixed effects)**

| Parameter | *b* | *SE* | *df* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| Intercept[a] | –1.29 | 0.04 | 99.16 | –32.15 | < .001 | 2.83 |
| Item type=word | –0.03 | 0.03 | 118.79 | –1.05 | .296 | 0.08 |
| Session=immediate | 0.16 | 0.02 | 87.72 | 8.94 | < .001 | 0.35 |
| Prime accuracy (Prime.ACC)=1 | 0.14 | 0.03 | 86.43 | 4.95 | < .001 | 0.30 |
| Prime length (Prime.nol)=7 | 0.07 | 0.02 | 77.58 | 3.72 | < .001 | 0.15 |
| Item type=word:Session=imm. | –0.11 | 0.01 | 5886.22 | –8.48 | < .001 | 0.25 |
| Item type=word:Prime.ACC=1 | –0.26 | 0.04 | 96.02 | –6.53 | < .001 | 0.56 |

*Note*. [a]Intercept levels: Item type = pseudoword, Session = delayed, Prime accuracy = 0, Prime length (Prime.nol) = 6.