Studies in Second Language Acquisition (2025), **47**, 417–439
doi:10.1017/S0272263124000706

REPLICATION STUDY

Optimizing distributed practice online

A conceptual replication of Cepeda et al. (2009)

John Rogers[1], Tatsuya Nakata[2] and Ming Ming Chiu[3]

[1]The Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong, Hong Kong; [2]College of Intercultural Communication, Rikkyo University, Toshima-ku, Tokyo, Japan and [3]The Education University of Hong Kong, Hong Kong, Hong Kong
Corresponding author: John Rogers; Email: john.rogers@polyu.edu.hk

(Received 05 May 2023; Revised 11 September 2024; Accepted 17 October 2024)

Abstract

This study conceptually replicates Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler's (2009, Experiment 1) study on the effects of distributed practice on second language (L2) vocabulary learning to examine its generalizability to a new context and population sample. The secondary focus of the paper is to examine the challenges and affordances of online data collection and participant recruitment sites. Both the original and our study examined the effects of distributed practice on two study sessions to learn L2 vocabulary assessed on a 10-day delayed posttest. Our results showed that the spaced conditions significantly outperformed the massed condition, mirroring the original study's findings. However, Cepeda et al.'s (2009) participants outscored our participants by 10–20% (in each experimental group) on the posttest. While these findings highlight the benefits of spacing towards learning and memory, they also underscore the challenges researchers may face when conducting experimental research in online environments.

Keywords: spacing effects; second language learning; data quality; web-based research; crowd-sourcing

Increasing the time between study sessions (*distributed practice* or *input spacing*) enhances learning and memory, as demonstrated in hundreds of verbal learning studies (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). In second language acquisition (SLA), vocabulary researchers have long recognized the utility of input spacing for theory, research, and pedagogy (e.g., Schmitt, 2000). Broader interest grew following discussions by Ellis (2006), Lightbown (2008), and a seminal empirical study by Bird (2010), which highlighted the potential benefits of distributed practice towards second language (L2) grammar development. However, subsequent empirical studies examining the effects of distributed practice on L2 outcomes have yielded mixed results (for syntheses, see Kim & Webb, 2022; Rogers, 2023; Serrano, 2022).

Various conceptual and methodological causes have been proposed to explain these equivocal findings. For example, differences in how studies have operationalized input spacing might account for some of the differences in the findings (Rogers, 2023). The majority of L2 learning studies have examined only two spacing gaps (e.g., one relatively short versus one relatively long). This impacts these studies' internal validity and the conclusions that we might draw from this research (Rogers, 2021). To elaborate, in a hypothetical study comparing two spacing gaps of 1-day and 4-day spacing intervals, if the 4-day spacing condition outperformed the 1-day condition, the researcher may conclude that longer spacing gaps were superior in this context. Changing the scenario, if there was a comparison with a 7-day spacing condition with a 14-day spacing condition, and the 7-day spacing condition outperformed the 14-day condition, then the conclusion may be the opposite, that shorter spacing gaps are superior. By contrast, if the study design had all four of the spacing conditions from above: a 1-day group, a 4-day gap, a 7-day group, and a 14-day group, then a fuller picture of the impact of spacing may emerge. These hypothetical examples illustrate how research designs that include only two spacing gaps can bias the results and hinder understanding of the impact of spacing on L2 performance and development. Finally, there have also been few replications of extant studies (Serrano, 2022). This hinders the external validity, or generalizability, of existing research.

To this end, as part of a larger, ongoing research project, the current study replicates Cepeda, Cepeda, Coburn, Rohrer, Wixted, Mozer, and Pashler (2009, Experiment 1), a highly cited study from the field of cognitive psychology that examined the learning of Swahili–English word pairs across six spacing gaps in a controlled laboratory setting. In contrast to the original experiment, we carried out the replication in an online environment using the Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) to create and host our experiment and recruited participants via Prolific Academic.

Background

First, we define our terms. Learning concentrated in a single, uninterrupted training session is *massed practice*. Learning spread across two or more training sessions is *distributed* or *spaced* practice. The superiority of spaced practice over massed practice is the *spacing effect.* The superiority of a longer spacing gap over a shorter spacing gap (e.g., a 1-week gap over a 1-day gap) is a *lag effect.* The gap between training sessions is the *intersession interval* (ISI). The gap between the final training session and the posttest session is the *retention interval* (RI; Cepeda et al., 2006).

Research into input spacing has a long history in the psychological sciences (for a review, see Wiseheart, Kim, Kapler, Foot-Seymour, & Kupper-Tetzel, 2019). More recent research has focused on the relationship between the intersession interval (ISI) and retention interval (RI). One such study is the focus of this replication, Cepeda et al.'s (2009, Experiment 1) study, which examined spacing and lag effects in the learning of Swahili–English word pairs. In their laboratory-based study, 215 English speakers were randomly assigned to one of six spacing conditions: massed (5-min gap), 1-day, 2-day, 4-day, 7-day, or 14-day ISI conditions. The treatments were conducted over two sessions, and the duration between the two treatment sessions varied according to the participants' assigned condition. For instance, in the 7-day interval condition, the second treatment session took place 7 days after the first session. In all six conditions, participants translated Swahili words to English on a meaning recall

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

posttest 10 days after the second training session (10-day RI). Their results showed that all five spaced conditions significantly outperformed the massed condition with large effect sizes (Cohen's $d > 1.0$). However, the spaced conditions showed no significant differences (i.e., 1-day ≈ 2-day ≈ 4-day ≈ 7-day ≈ 14-day > massed). Hence, this study evidenced a spacing effect but not a lag effect.

In another study, Cepeda, Vul, Rohrer, Wixted, J& Pashler (2008) conducted a large-scale, online learning experiment to further explore the ISI–RI relationship; 1,350 participants recruited online learned some trivia facts (e.g., "What European nation consumes the most spicy Mexican food? – Answer: Norway," p. 1097). Participants were assigned randomly to one of 26 experimental conditions, which comprised 12 ISIs (0 to 105 days) and 4 RIs (7 to 350 days). Their results showed a nonmonotonic, inverted U-shaped relationship between the ISI–RI ratio and retention. Specifically, posttest results suggested that the optimal spacing interval was approximately 10% to 40% of the RI. For example, for a posttest given 30 days after the last training session, an ISI of 3 days (30 days × 10%) to 12 days (30 days × 40%) yielded the highest retention. When spacing intervals fell short of the optimal ISI, increasing spacing increased retention, possibly because longer spacing induced retrieval difficulty that aided learning (*desirable difficulty*). When the spacing intervals exceeded the optimal ISI, increasing spacing decreased retention. Their study also suggested that the optimal ISI–RI ratio decreased as the RI increased, 20% to 40% for the 1-week RI and 5% to 10% for the 350-day RI though the results differed across assessment tasks (e.g., recall vs. multiple choice). Experimental conditions within the optimal ISI–RI range showed 10% to 111% improvement over ISI–RI conditions outside of the optimal range, with medium-sized ($d = .6$) to large-sized effects ($d = 1.7$).

Effects of intersession spacing on L2 vocabulary learning

Studies such as Cepeda et al. (2009; also Cepeda et al., 2008) that have provided evidence of the ISI–RI relationship have been hugely influential across the psychological sciences. This influence extends to the field of SLA, where researchers have sought to apply these optimal ISI–RI spacing models to L2 learning (e.g., Bird, 2010; Rogers, 2015; Suzuki & DeKeyser, 2017). However, empirical studies have shown mixed results (Rogers, 2021; Serrano, 2022). For example, in L2 vocabulary studies, some studies have found no significant difference between ISI conditions (Rogers & Cheung, 2021). Another study (Serrano & Huang, 2018) indicated that the long-spaced condition may yield greater long-term retention relative to the short-spaced condition. Three have shown the advantage of shorter ISIs over longer ISIs (Küpper-Tetzel, Erdfelder, & Dickhäuser, 2014; Rogers & Cheung, 2020; Serrano & Huang, 2023). These conflicting results may be due to the narrow range of ISI–RI ratios examined as part of their experimental designs (Rogers, 2021). As a result of this piecemeal approach (i.e., examining only a limited range of ISI–RI combinations in each study), it is unclear to what extent these findings might simply be artifacts of their experimental designs.

Despite being a mature area of inquiry within SLA, spacing research has several limitations. Methodological limitations include the limited number of ISI conditions and the narrow range of ISI–RI ratios within individual studies. Among L2 vocabulary studies, Cepeda et al. (2009, Experiment 1) examined the most ISI conditions (six) and the widest range of ISI–RI ratios (.03% to 140%). As such, the results of their study provide the clearest interpretation to date, have been highly influential, and warrant replication to examine their generalizability.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

Replication research

Replication is an empirical, methodological approach that can consolidate and advance scientific understanding, so it is vital for the credibility and growth of a discipline (Porte & McManus, 2019). Replication studies repeat a previous study, copy its design and methods (with or without changes), collect new data, and systematically compare the new results with those of the previous study. Doing so provides a systematic framework for better understanding both the methodology and results of previous research and a basis for an understanding of a theoretical construct in different experimental scenarios. There are four approaches to replication: exact, close, approximate, or conceptual replications, depending on the number of manipulated variables (Porte & MacManus, 2019). An exact replication follows the previous study's entire design without any changes. Close or approximate replications modify one or two major variables, respectively. Conceptual replications change more than two major variables from the original study. The current study is an example of the latter in that it sets out to conceptually replicate Cepeda et al.'s (2009) study on the effects of spacing on the learning of L2 vocabulary.

Validity of online data collection

Researchers across the social sciences, including SLA, have increasingly turned to online experimental platforms for data collection. Technological advancements, the development of specialized online platforms, and necessity in many contexts (e.g., the COVID-19 pandemic) have driven this shift (Uittenhove, Jeanneret, & Vergauwe, 2023). With regard to online experimentation, it is important to differentiate between platforms for hosting experiments (i.e., data collection) and platforms for participant recruitment. As different specialized platforms have been developed for each purpose, they have implications for data quality, as discussed below.

Examples of dedicated platforms for hosting experiments online include *Gorilla Experiment Builder* and *PsychoPy* (Peirce et al., 2019; see Mathôt & March, 2022 for a discussion of other platforms). These platforms allow researchers to build and host experiments and collect data but do not recruit participants. The researcher must direct the participants to the platform (Mathôt & March, 2022; Rodd, 2024).

To recruit participants, researchers can use conventional methods (e.g., place posters on university campuses and direct participants to the experimental platform, now often via a link or quick response [QR] code). Alternatively, researchers can crowdsource participants for their experiments via dedicated participant recruitment platforms. The most well-known platforms are Amazon Mechanical Turk (*Mturk*) and Prolific Academic (*Prolific*; see Mathôt & March 2022 for an extensive list). There are several advantages to crowdsourcing participants for online experiments. For instance, *Mturk* and *Prolific* allow (a) access to a large, diverse participant pool, (b) researchers to screen participants across many criteria, and (c) quick and remote data collection (Rodd, 2024). Notably, access to large participant pools can help address issues of low statistical power. There is a growing awareness of underpowered quantitative research across the psychological sciences and SLA. Underpowered research can lead to sampling bias and results that cannot be replicated by other researchers (see Andringa & Godfroid, 2023; Rodd, 2024, for discussions).

Despite the advantages of collecting data online, researchers have also recognized some challenges to preserving data quality. For example, unlike in face-to-face data collection, online experimenters do not directly oversee data collection processes, which may increase participant distractions/inattentiveness and fraudulent behaviors (Newman, Bavik, Mount, & Shao, 2021; Uittenhove et al., 2023). To reduce them, it is

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

recommended to use (a) fair compensation for participants, (b) items to check participant attention, (c) explicit instructions, questions, and warnings, (d) bot check items to safeguard against fraudulent behavior (Mathôt & March, 2022; Newman et al., 2021), (e) email/text reminders for multisession/longitudinal studies, (f) screening data for "surprisingly good" performance, and (g) benchmarking online data against face-to-face data (Cepeda et al., 2008; Rodd, 2024). See the Methods section for a discussion of safeguards utilized in the current study.

Overall, validation studies have shown that (a) behavioral experimental data collected online can be of good quality (Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2021; Hartshorne, de Leeuw, Goodman, Jennings, & O'Donnell, 2019; Uittenhove et al., 2023), including L2 data (Kim, Liu, Isbell, & Chen, 2024; Nagle & Rehman, 2021; Patterson & Nicklin, 2023; Ruiz, Chen, Rebuschat, & Meurers, 2019), and (b) online participants often outperform face-to-face participants on quality control measures (e.g., attention, see reviews by Hartshorne et al., 2019 & Newman et al., 2021). In fact, in their comprehensive review of the behavioral literature, Hartshorne et al. concluded that "internet volunteers comply with instructions and answer truthfully at rates matching or exceeding lab-based subjects, resulting in data with similar psychometric validity" (2019, Appendix C; see also Rodd, 2024).

In sum, there is evidence that online experimental research can be carried out validly and reliably. However, this evidence should not be construed to indicate that *all* online experiments will necessarily be valid and reliable (Rodd, 2024). For example, at present, technology limits the transfer of some experimental paradigms to online settings. Examples include paradigms that utilize eye tracking (e.g., the visual world paradigm), where efforts are ongoing to validate its use for data collection in online environments (see, e.g., Van der Cruyssen et al., 2024). Finally, collecting data online involves some special considerations related to technology, recruitment, and participant performance. For a full discussion of these issues, specific recommendations, and checklists for researchers conducting behavioral research online, see Rodd (2024).

To our knowledge, in the field of SLA, only one multisession, experimental study of distributed practice that recruited participants via crowdsourcing platforms has been published to date. Serfaty and Serrano (2024) examined the role of practice, specifically the impact of relearning sessions, on the learning of L2 grammar. In their study, 122 adults aged between 18 and 30 years, fluent in English from a variety of countries, were recruited via Prolific Academic. In the initial training session, participants were presented with the target structure, an artificial language. In the second phase, participants had to learn 12 sentences to criterion. Following this, participants completed either one, two, three, or four additional relearning sessions in which they practiced the same 12 sentences. A 1-day gap separated adjacent sessions, which was chosen to minimize attrition. Fourteen days after the last training session, participants completed a delayed posttest. As noted by the authors, the timing of the 14-day delayed posttest was partly chosen due to restrictions of the Prolific payment system, as it was the longest delay that would be possible without the Prolific system automatically paying participants. Delayed posttest results indicated that participants who completed three or more relearning sessions scored higher on productive measures of learning. Of importance and methodological relevance to our current project, no attrition of participants was reported.

This study

This online experiment replicates Cepeda et al.'s (2009, Experiment 1) face-to-face laboratory study of input spacing on L2 vocabulary learning to assess the extent to

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

which their findings generalize across learning contexts. As noted above, the limited number of spacing conditions in most L2 studies impedes the interpretation of their results. Cepeda et al.'s (2009) study has examined the most ISI conditions and the widest range of ISI–RI ratios of L2 vocabulary studies. Hence, the replication of Cepeda et al. (2009) will help determine its generalizability and contribute another multi-ISI (>2) experiment to the L2 spacing literature.

Furthermore, we examine the viability and validity of current online experimental procedures. Researchers have increasingly turned to online platforms for experimental studies. Only one multisession experimental study (Serfaty & Serrano, 2024) using similar experimental platforms and recruitment procedures (e.g., Prolific Academic) has been published to date. Given the challenges of online data collection, this study will document the challenges and considerations of online research.

Analyses of participants' training performance and posttest performance are crucial to our study. Using these data, we examine the link between behaviors during the learning process (trials required to reach criterion during the training phase) and learning outcome (posttest performance). As identical data were available from Cepeda et al.'s (2009) study, we used them to benchmark our participants' training performance and posttest performance which helped validate our experimental procedure.

Our study addresses two research questions: To what extent does spacing (massed, 1-day, 2-day, 4-day, 7-day, 14-day intersession interval) influence the learning of L2 vocabulary learning as measured on a 10-day delayed meaning-recall posttest? To what degree do these results align with the results of the original study? The second focus of the paper is to examine the challenges and affordances of online data collection and participant recruitment sites.

Methods

Participants

Whereas Cepeda et al. (2009) had English-speaking participants from the United States, we recruited United Kingdom participants via Prolific Academic (https://www.prolific.co/). Participation in this experiment was limited to UK participants (born and currently located in the UK) for theoretical and practical reasons. First, we were interested in examining the generalizability of Cepeda et al.'s (2009) findings to a different English as first language (L1) context. Unlike other online studies that have sampled participants without any geographical restrictions in place, our geographical restriction ensured that our sample of L1 speakers differed from those in the original study. Second, this study was designed as part of a larger project sampling from the UK.

Like the original study, our participants were L1 speakers of English. However, participants in the original study were undergraduate students at the University of California, San Diego, whereas our study required participants to only have a minimum of secondary schooling (or equivalent). All participants gave informed consent before joining the study. They received £12 (approximately $16) for completing the study: £6 for completing the first of the three experimental sessions and £6 for completing the final (third) session. Due to Prolific's policy, participants received two separate payments (rather than one payment at the end; See also Discussion section).

To approximate the original study's 215 participants, we recruited 569 participants, but only 222 (39%; see Table 1) completed the experiment (101 males; 117 females; 4 nonbinary; *Mean* age = 31.5, *SD* = 12.4).

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

Table 1. Attrition by experiment condition, session, and stage of study

Session	Stage	Massed	1-day	2-day	4-day	7-day	14-day	Total
Training 1	Consent	81	88	100	81	98	130	569
	Demo 1	78	85	99	81	96	130	569
	Demo 2	77	85	99	81	95	130	567
	Pretest	77	84	99	81	94	130	565
	Practice	75	84	96	80	92	128	555
	Presentation	75	83	96	80	92	127	553
	Retrieval 1	74	82	96	79	91	124	546
	Bot check 1	69	69	85	70	81	115	489
	Delay 1	69	69	85	70	79	111	483
Training 2	Retrieval 2	66	47	53	40	40	45	291
	Bot check 2	66	45	52	38	37	40	278
	Delay 2	66	45	52	38	36	39	276
Posttest	Word-form recognition	47	40	45	31	31	33	225
	Meaning recall	46	39	45	31	31	33	222
	Meaning recognition	46	39	45	31	31	33	222
	Debriefing	46	39	45	31	31	33	222
	Finish	46	39	45	31	31	30	222
	Completion %	57%	44%	45%	32%	32%	23%	39%

Note: The numbers indicate the number of participants who completed each stage of the experiment.

Table 2. Comparison of the methodological features of Cepeda et al. (2009, Experiment 1) and the present study

	Cepeda et al. (2009) Experiment 1	Present study
Location	Laboratory	Online (via Gorilla)
Participants	215 undergraduate students from UC San Diego.	A total of 222 adult participants recruited via Prolific. UK-based. English L1, no background in Swahili or similar languages. Sixty participants did not meet our demographic requirements or failed validity checks, so we analyzed data from 162 participants.
ISI–RI manipulations	6 ISI–RI combinations	Identical
Materials	40 L2 (Swahili)–L1 (English) word pairs	Identical
Treatment	• Presentation of target items • Meaning recall practice (with feedback)	Identical
Number of retrieval trials during treatment	Training Session 1 • Practiced to the criterion of two correct answers Training Session 2 • Practiced twice	Identical
Intersession intervals (ISIs)	0, 1, 2, 4, 7, and 14 days	Identical
Retention interval (RI)	10 days	Identical
Pretest	None	Meaning recall test
Dependent variables (process)	None reported	• Performance during Training 1 (response accuracy) • Performance during Training 2 (response accuracy and latency)

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

Table 3. Safeguards for data quality incorporated in the current study

Recommended practice	Implementation in the current study
Specify experiment-specific data quality concerns (see Rodd, 2024)	We brainstormed potential threats to data quality (Supplementary Appendix S9). This list was later used in the data screening process.
Fair compensation for participants	Participants received £12 for the study. This amount is considered fair, taking into consideration the length and nature of the study. It was in line with Prolific's payment guidelines.
Attention check items	Each experimental session included two attention check items randomly interspersed amongst target items.
Explicit instructions, questions, and warnings	Participants were given explicit instructions and reminders, such as reminders when they should return for the next stage of the experiment, reminders about expectations of performance, and so forth. All tasks included two practice items with feedback. These were included to help ensure participants understood task requirements.
Bot check items	Two bot check items were included at different points in the experiment (Training Phase 1 & 2).
Email/text reminders in longitudinal studies	Messaging and email reminders were sent via the Prolific messaging system.
Screening for surprisingly good performance	Data were screened for suspicious performance: 1. Training phase 2. Testing phase
Benchmarking online data against face-to-face data	Our online data were benchmarked against Cepeda et al.'s (2009) data collected in face-to-face laboratory environment.

Although Prolific Academic prescreened participants by L1, four participants reported that English was not their L1 in our survey conducted at the beginning of the first training session, so we excluded them. Because our study involved the learning of Swahili words (for details, see Materials below), only those without familiarity with Swahili or similar languages (e.g., Arabic, Bantu) were invited to participate. As 30 participants reported familiarity with Arabic (16), a Bantu language (13), or both (1) on our pretraining survey, we excluded them from our analyses, leaving 188 participants. Further data screening excluded 26 more participants (see Validity section below). As a result, we only analyzed data from 162 participants.

Materials

We used Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Following Cepeda et al.'s (2009) Experiment 1, we used 40 L2 (Swahili) to L1 (English) word pairs (e.g., *samaki-fish*; for the full list, see Appendix S1 in online supplementary materials; all study materials and data are publicly available at iris-database.org).

Procedure

We followed Cepeda et al.'s (2009) Experiment 1 procedure unless otherwise noted (Table 2). All participants were asked to join three experimental sessions: Training 1, Training 2, and Posttest (see Figure 1).

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

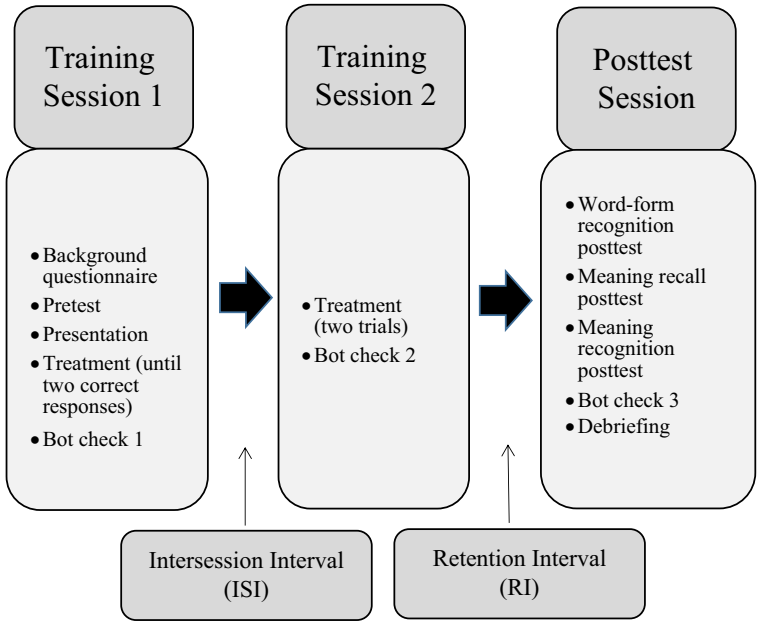

Figure 1. Experimental procedure.

Training session 1

Participants completed a short survey (gender, age, L1, L2s, any familiarity with Swahili or related languages) and a meaning recall pretest. In the pretest, each target Swahili test item was presented individually and randomized across participants. Participants were prompted to type its English translation/definition into an onscreen box. This pretest helps identify participants with prior knowledge of the target Swahili words. As a meaning recall test is relatively challenging, its results may miss some partial knowledge (Laufer & Goldstein, 2004). Still, we choose the meaning recall pretest to be identical to the meaning recall posttest (also used in Cepeda et al., 2009), thus allowing for direct comparison.

After the pretest, participants practiced learning Swahili words using two practice items, followed by the treatment. The treatment had a presentation phase and a retrieval phase. In the presentation phase, participants were presented with 40 Swahili–English word pairs, one at a time (e.g., *samaki-fish*) for 7 s each, randomized across participants (see Figure 2).

Next, in the retrieval phase, participants were presented with each Swahili word (e.g., *samaki*) one at a time and were prompted to type its English translation into an onscreen box (no time limit, see Figure 3). After each response, participants received feedback visually and auditorily (*Correct!* or *Incorrect!*), followed by the screen display of the correct answer for 5 s. Target items were repeated in a random order in blocks of 40 items until each item was answered correctly twice. Items correctly answered twice were removed. Item order was randomized and manipulated to ensure a minimum of 20 items between repetitions of any item. When 20 words or fewer remained, maximum separation was maintained between appearances of the same word.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

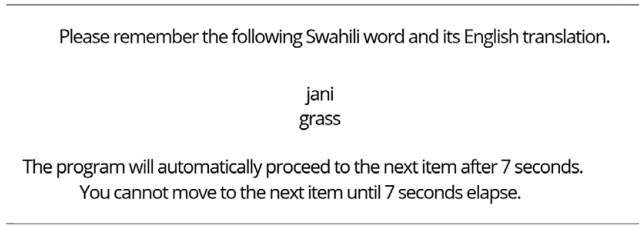

Figure 2. Example from the treatment (presentation phase).

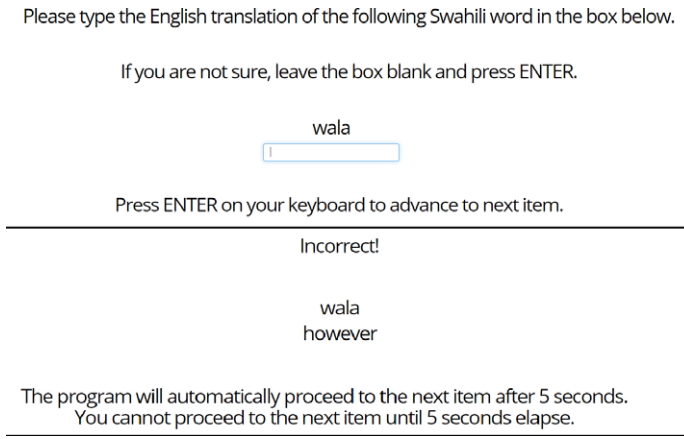

Figure 3. Example from the treatment (retrieval phase).

Participants then completed a bot check (e.g., type the name of the animal in the picture, e.g., *cat*, *rabbit*). As bots often respond with random words to such open, straightforward questions, the resulting answers are typically nonsensical (Newman et al., 2021), or they are unable to complete the question (or captcha). Failing to respond to a bot check meant that a participant would not be able to continue in the experiment.

Training session 2

Training Session 2 only had the retrieval phase (no presentation phase). Participants practiced the randomized list of items in a meaning recall format twice, regardless of learner performance.

Posttest session

The participants completed three posttests in this order: (1) word-form recognition, (2) meaning recall, and (3) meaning recognition. Although Cepeda et al.'s (2009) original study only had a meaning recall test, we added a word-form recognition test and a meaning recognition test to detect lower levels of partial or incomplete L2 word knowledge (Waring & Takaki, 2003). See Appendix S2 in online supplementary materials for descriptions of word-form recognition and meaning recognition posttests. The formats of the meaning recall pretests and posttests were identical.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

After the posttests, participants were asked if they encountered any of the target vocabulary items outside this study, and, if so, to elaborate (e.g., "I found some word meanings in a dictionary").

Validity

Validity threats include attrition, L2 knowledge, inattention, bots, external resources, and actual vs. assigned ISI or RI. (We generally preserved data, excluding only clearly compromised data.)

Attrition

As noted, one of the greatest challenges to our study was participant attrition. Only 39% of the participants completed the study (see Table 1). Longer ISI often reduced completion rates (massed: 57% > 14-day ISI: 23%). Among the 347 participants who dropped out, most did so between experimental sessions (206/347 = 59%), and many between the first and second training sessions (155/347 = 45%). Regardless of whether participants completed the experiment or not, their performance was similar on pretest accuracy (1.1% vs. 2.6%), trials to criterion in Training Session 1 (164.8 vs. 167.1), and response accuracy in Training Session 2 (70.3% vs. 67.9%).[1]

L2

On the pretest, two participants correctly answered many questions (14 and 15 out of 40), so they were excluded. Another 50 participants answered at least one item correctly on the pretest (M = 1.4, max = 6), so their correct responses served as covariates in subsequent data analyses.

In Training Session 1, participants learned the criterion of two correct recalls. Cepeda et al.'s (2009) participants required 231.4 trials on average to do so (SD = 62.2; range: 134–512). Many studies identify and omit outliers 2 or more SDs from the mean (Jiang, 2012), 107 trials for the original study (107 = 231 − [2 × 62]). Hence, we excluded the 25 participants who learned all the words in less than 107 trials. Compared with the included participants, these excluded participants had higher posttest score means (70.0% > 54.7%).

Attention check items

Each of our three sessions included two attention check items (e.g., please type *monkey* into the box and press <enter>) among the target stimuli. This helps ensure data quality as some participants might thoughtlessly click through items.

Two participants failed both attention check items in Training 1, so we excluded these two participants. No participants missed both items in the other sessions. Some participants missed one attention check: 10 participants during Training Session 1; two during Training Session 2; and one during the Posttest Session. One of the participants

[1]The pretest accuracy rate for incomplete participants was inflated due to an outlying value, specifically a participant who scored 100% (40/40 accurate answers). If excluded from the analysis, the pretest average of incomplete participants is 1.5% [1.2%, 2.9%].

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

who failed one attention check in Training Session 2 was excluded for another reason (few trials to criterion).

The participants who failed one attention check mostly failed only one throughout the entire experiment. For example, the 10 participants who failed an attention check in Training Session 1 did not fail another attention check throughout the remainder of the experiment. Hence, we interpret their missed attention checks as slips rather than systematic inattentiveness.

Bot check items

Eight participants incorrectly answered the bot checks and were excluded from the analyses.

External resources

At the end of the experiment, four participants acknowledged using online dictionaries or other resources during the experiment. Hence, they were excluded from the analyses.

Actual vs. assigned ISI

Participants did not always complete their second experimental sessions according to schedule. Following Cepeda et al. (2009) original study, our strategy was to code participants' ISI categorically. To do so accurately, our strategy when participants did not complete the experiment on schedule was to round them to the nearest ISI condition. For example, if a participant in the 2-day condition returned after 3.1 days, we recoded them to belong to the 4-day ISI group. However, participants who returned after 2.9 days remained in the 2-day group. We recoded 15 participants for ISI.

Actual vs. assigned RI

Most participants adhered to the 10-day RI for the posttest (M = 10.5, SD = 1.2, range: 9.0–17.3; see Appendix S3 in online supplementary material). Analysis of variance (ANOVA) showed no significant differences between the six experimental groups in their RI, F (5, 156) = .72, p = .613. Although three participants' RIs exceeded 16 days, their performance otherwise appeared normal, so we retained them in our analyses.

Based on the above criteria, we excluded 60 (including 34 participants due to L1) participants. So, we analyzed data from 162 participants.

Table 3 presents a summary of recommended data-safeguarding practices (e.g., Rodd, 2024) and how we implemented them in the current study.

Scoring

Gorilla automatically scored each item on the pretest and meaning recall posttest as correct or incorrect (1 vs. 0). One author and a research assistant independently double-marked 25% of all responses across participants. Following Cepeda et al. (2009), misspelled words (e.g., "poket" instead of "pocket") were marked correct. As synonyms ("road" instead of "street") reflect knowledge of the target forms (i.e., participants wrote the correct meaning, just not the exact meaning that the researcher had in mind), they

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

were marked correct. Following Cepeda et al. (2009), we used a double-marking scheme, where all coding was blind to experimental conditions. Interrater reliability was high (Krippendorf's α = .99). The research assistant independently marked the remaining 75% of responses. Among the included participants (n = 162), 14 participants correctly used 16 synonyms on the pretest, and 25 participants correctly used 32 synonyms on the posttest.

Data analysis

We analyzed our data using (a) analysis of variance (ANOVA) and covariance (ANCOVA) via JASP Version 0.16.3 (2022), and (b) *mixed-effects analyses* (aka *multilevel analyses*, Hox et al., 2017) via MLwiN 3.05 (Charlton, Rasbash, Browne, Healy, & Cameron, 2020).[2] Our ANOVAs aid comparability with Cepeda et al.'s (2009) original study, while our mixed-effects analyses make fewer assumptions and allow for more accurate modeling of the data. Like Cepeda et al. (2009), we used ANOVA to compare the posttest scores across the experimental conditions. We followed this with an analysis of covariance (ANCOVA) where pretest scores, trials to criterion in Training Session 1, and performance accuracy of Training Session 2 were covariates. Then, to more accurately analyze the data, we used *mixed-effects analyses* (aka *multilevel analysis* or *hierarchical linear modeling*; see Appendix S4 in online supplementary materials for details). We ran two mixed-effects analyses. These mirror the structure of the ANOVA and ANCOVA. Like the ANOVA, the first mixed-effects analysis includes no control variables (covariates). Identical to the ANCOVA, our second mixed-effects analysis includes pretest scores, trials to criterion in Training Session 1, and performance accuracy in Training Session 2 as control variables. For these mixed-effects analyses, spacing (massed, 1-, 2-, 4-, 7-, vs. 14-day) served as a fixed effect (Models 1 and 2). Participant and item served as random effects (Models 1 and 2). Pretest scores, trials to criterion in Training Session 1, and performance accuracy of Training Session 2 served as additional control variables (random effects) in Model 2.

Statistical power

Statistical power differed across levels. For α = .05 and a past effect size of .22 (Cepeda, 2009), the statistical power for 162 participants was .80, and for the 6,480 presented words (162 participants × 40 words = 6,480 posttest item responses) exceeded .99.

Results

Training sessions 1 and 2

The number of trials required to reach the criterion of two correct answers in Training Session 1 did not differ significantly across the six experimental groups, F (5, 156) = .69, p = .632, η^2 = .02, suggesting that their L2 vocabulary learning capabilities did not differ.

[2]MLwiN is an alternative to *R* for running mixed-effects models (see, e.g., Quené & van den Bergh, 2008 for a discussion). Whereas *R* is able to run a broad range of statistical procedures via different packages, MLwiN is a specialized multilevel modeling software package, specifically designed for the type of analysis (multilevel) that we carry out in this study.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

ANCOVA of performance accuracy in Training Session 2 with covariates (pretest score, accuracy rate during Training Session 2, and Training 1 trials to criterion) showed a significant group effect: F (5, 153) = 18.46, p < .001, η^2 = .32 (for descriptive statistics, see Appendix S5 in online supplementary materials). Post hoc Tukey tests revealed that the massed participants outscored the 2-, 4-, 7-, and 14-day spaced participants (Cohen's ds = 1.10 [.56, 1.64], 1.42 [.88, 1.95], 1.93 [1.33, 2.52], and 2.31 [1.71, 2.91], respectively; numbers in brackets indicate the 95% confidence interval). Also, the 1-day ISI participants outscored the 7- and 14-day ISI participants (d = 1.26 [.67, 1.85] and 1.64 [1.05, 2.24], respectively). The 2-day ISI participants outscored the 14-day ISI participants (d = 1.21 [.64, 1.79]). No other pairwise difference was statistically significant. Overall, these results showed that shorter ISI conditions yielded greater Training Session 2 accuracy; possibly, the smaller duration gaps between Training Sessions 1 and 2 reduced memory decay.

Posttest session

We only discuss the results of the meaning recall (translation) posttest in detail because (a) only this test is in both the original study and this study, and (b) high scores for the word-form recognition and meaning recognition posttests suggest ceiling effects (for descriptive and inferential statistics of the word-form and meaning recognition posttests, see Appendix S6 in the online supplementary materials). Meaning recall posttest reliability (Cronbach's α = .83) was acceptable. See descriptive statistics for the original study and the current study in Table 4 and Figure 4.

ANOVA and ANCOVA

Our initial ANOVA results showed significant differences among groups in their meaning recall posttest scores, F (5, 155) = 2.89, p = .016, η^2 = .08. Tukey's multiple comparison test showed that the 2-day group significantly outperformed the massed group (p = .009, d = .91). No other differences were statistically significant ($p \geq .067$, $.05 \leq d \leq .71$).

Next, we ran an ANCOVA, controlling for pretest score, Training Session 1 performance (number of trials to criterion), and Training Session 2 performance (accuracy

Table 4. Proportion of correct response (%) on the meaning recall posttest in the current study, compared with performance on Cepeda et al. (2009)

	Present study			Cepeda et al. (2009)		
	n	*M*	*SD*	*n*	*M*	*SD*
Massed	33	43.3 [35.7, 50.9]	21.5	31	54.9 [47.2, 62.6]	21.0
1-day	26	57.7 [47.5, 67.9]	25.3	31	73.9 [67.4, 80.5]	17.9
2-day	27	62.7 [53.5, 71.9]	23.2	30	68.8 [63.0, 74.7]	15.8
4-day	29	58.2 [49.3, 67.1]	23.3	29	68.5 [61.4, 75.5]	18.5
7-day	23	56.2 [46.8, 65.6]	21.7	29	69.4 [62.9, 75.9]	17.1
14-day	24	52.3 [42.8, 61.8]	22.4	32	65.5 [58.4, 72.5]	19.6

Note: [ ] indicates a 95% confidence interval for *M*.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

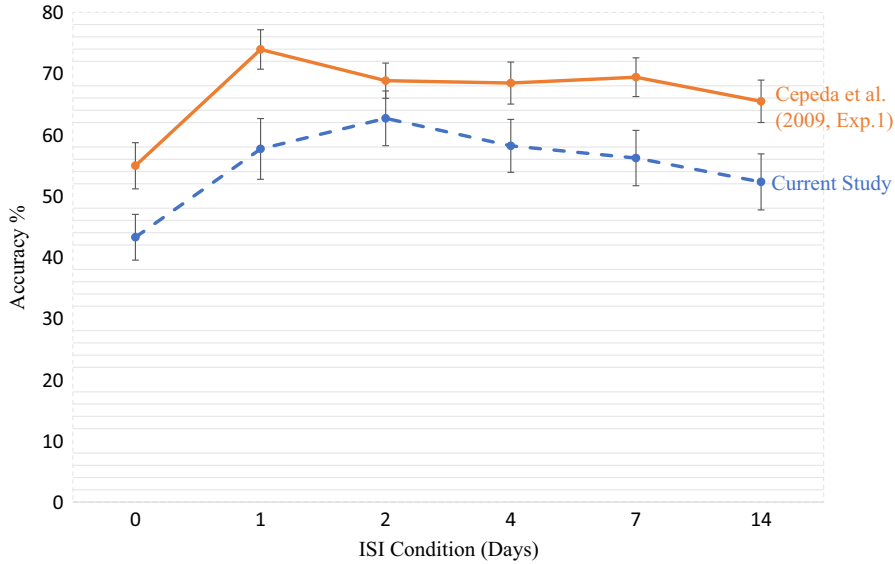

Figure 4. Comparison of posttest performance in the current study and Cepeda et al. (2009, Experiment 1).

rate); meaning recall posttest scores still differed across groups, F (5, 153) = 23.97, $p < .001$, $\eta^2 = .28$. All five spaced groups significantly outperformed the massed group (spacing effect; see Tukey's multiple comparison test results in Table 5). Also, the 4-day, 7-day, and 14-day groups outperformed the 1-day group. No other comparisons were statistically significant.

Mixed-effects modeling

The first mixed-effects analysis (no control variables) accounted for only 2.0% of the differences in *correct* responses (see Appendix S7 in the online supplementary file). The second mixed-effects analysis with control variables showed that training and spacing were linked to correct answers. Fewer *training 1 trials to criterion* and greater *training 2 accuracy* were both linked to a greater likelihood of a correct answer on the posttest (respectively by −.1% and +39%; see Model 3 of Table 6). *Training 2 accuracy* mediated 47% of the link between *training 1 trials to criterion* and *correct* ($z = 3.68$, $p < .001$). Together, these variables accounted for 8% of the variance in *correct* responses.

Furthermore, participants in the *1-day*, *2-day*, *4-day*, *7-day*, and *14-day* conditions outperformed participants in the *0-day* condition, controlling for other explanatory variables (respectively by +27%, +35%, + 38%, +38%, and +40%; see Model 4 of Table 5 and Figure 5), accounting for an additional 9% of the variance. All other variables (e.g., *pretest score*) were not significantly related to correct responses. Overall, these results (with control variables) accounted for 17% of the variance in *correct* (far more than the model without control variables: 17% > 2%).[3]

[3]Following a suggestion from an anonymous reviewer, we conducted a follow-up analysis with time-on-task as a covariate. This did not significantly impact the results of the study in terms of statistical significance or effect size. Results of the follow-up analysis are presented in Appendix S8 in the online supplementary file.

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

Table 5. Results of Tukey's multiple comparison test for the meaning recall posttest scores.

	1-day		2-day		4-day		7-day		14-day	
	p	*d*	*p*	*d*	*p*	*d*	*p*	*d*	*p*	*d*
Massed	< .001	1.63 [.78, 2.48]	< .001	2.28 [1.36, 3.20]	< .001	2.53 [1.58, 3.47]	< .001	2.73 [1.68, 3.78]	< .001	2.99 [1.88, 4.10]
1-day			.183	.65 [–.19, 1.49]	.019	.90 [.06, 1.74]	.005	1.11 [.18, 2.04]	< .001	1.37 [0.40, 2.33]
2-day					.945	.25 [–.57, 1.06]	.639	.45 [–.43, 1.33]	.169	0.71 [–0.19, 1.61]
4-day							.978	.21 [–.64, 1.06]	.577	0.47 [–0.39, 1.32]
7-day									.952	0.26 [–0.62, 1.14]

Note: [ ] indicate a 95% confidence interval for *d*.

Table 6. Summary of mixed-effects analysis of correct on the meaning recall posttest (with control variables).

	Model 1				Model 2			
Explanatory variable	Coefficient	*SE*	*p*	Odds Ratio	Coefficient	*SE*	*p*	Odds Ratio
Pretest score	6.08	–7.23	.401		5.83	–6.62	.379	
Training 1 trials to criterion					–.01	.00	< .001	1.00
Training 2 accuracy								
1-day								
2-day								
4-day								
7-day								
14-day								
Constant	.23	–.10	.016	1.06	.23	–.09	.011	1.06
Variance at each level								
Participant (29%)	.00				.12			
Word (71%)	.00				.00			
Total variance explained	.00				.03			

	Model 3				Model 4			
Explanatory variable	Coefficient	*SE*	*p*	Odds Ratio	Coefficient	*SE*	*p*	Odds Ratio
Pretest score	5.90	–6.24	.345		5.72	–5.83	.327	
Training 1 trials to criterion	–.003	.00	.003	1.00	.00	.00	1.000	
Training 2 accuracy	2.13	–.31	< .001	1.39	4.30	–.38	< .001	1.49
1-day					1.19	–.21	< .001	1.27
2-day					1.74	–.22	< .001	1.35
4-day					1.96	–.23	< .001	1.38
7-day					2.06	–.25	< .001	1.39
14-day					2.24	–.27	< .001	1.40
Constant	.22	–.08	.006	1.06	.30	–.06	< .001	1.07
Variance at each level								
Participant (29%)	.28				.60			
Word (71%)	.00				.00			
Total variance explained	.08				.17			

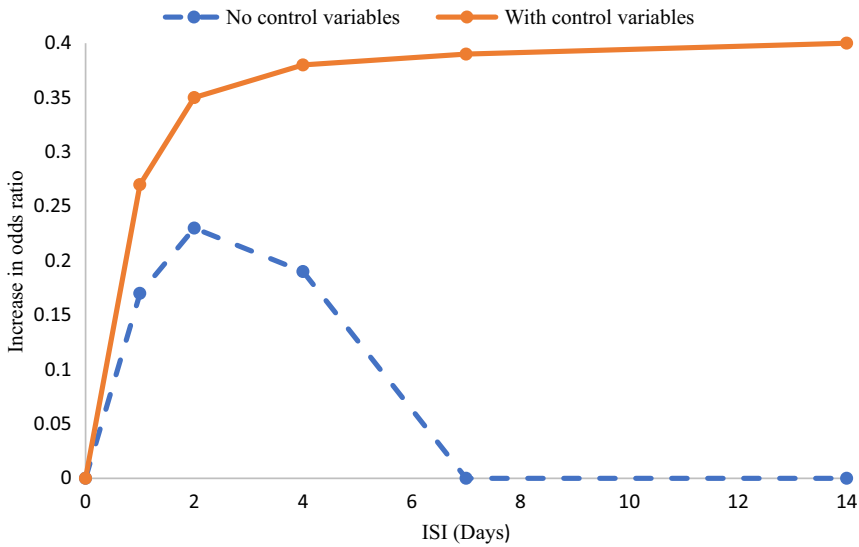

Figure 5. Additional benefits of spacing.

Discussion

This online study replicated Cepeda et al. (2009, Experiment 1), which examined six different spacing conditions (massed, 1-, 2-, 4-, 7-, and 14-day ISIs) on the learning of Swahili–English word pairs, as measured following a 10-day delay. Cepeda et al.'s (2009) Experiment 1 showed a significant difference between the spaced conditions and the massed condition (spacing effect) but not between spacing conditions (no lag effects). Our results support the original findings from Cepeda et al. (2009, Experiment 1) and demonstrate that these extend to a new population sample and online environment. On a methodological level, this study is one of the first online, multisession experimental studies in our field and highlights some challenges of carrying out such research.

Our study's results differed by analyses and controlled variables. ANOVA and ANCOVA showed the following: On the ANOVA, only the 2-day ISI posttest scores exceeded those of the massed condition. On the ANCOVA, which controlled for pretest scores, Training Session 1 trials to criterion, and Training Session 2 accuracy rate, the posttest scores of all five spacing conditions far exceeded those of the massed condition (spacing effects; ds > 1.5). Also, the 4-, 7-, and 14-day groups significantly outperformed the 1-day group (lag effects; ds > .9).

According to the mixed-effects analysis results, posttest scores of the spaced conditions exceeded those of the massed condition, and Training Session 2 accuracy was linked to posttest performance. Modeling control variables accounted for much more variance than omitting them (17% > 2%); this reduces *omitted variable bias*, thus supporting their inclusion in our analyses. Also, correct (vs. incorrect) responses differed mostly across words (71%) rather than across participants (29%). Intralexical factors, such as word length, pronounceability, orthography, and semantic features (e.g., abstractness, polysemy, etc.; see Laufer, 2012; Peters, 2016 for discussions) might have contributed to this item variance. As our study only aimed to validate our online experiment procedure, we used the same item set as Cepeda et al.'s (2009) original

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

study, and we did not control for these variables. Although the order of the items was fully randomized by participant and by stage of the experiment, such factors might have influenced the findings. These findings highlight the importance of controlling for these variables at the design stage of the experiment and when analyzing performance at the item level (rather than only the participant level, e.g., ANCOVA).

Although these results align with those of Cepeda et al.'s (2009) original study, our study required covariates that were not controlled in the original study. It is worth discussing some possibilities as to why this might be the case. For instance, our study participants are from a different population (US vs. UK) and participated in a different context (supervised face-to-face vs. unsupervised online) compared with the original study. These differences may have influenced the results. To elaborate, regarding the online context, the lower degree of experimental control present in the online environment, relative to face-to-face laboratory settings, might have reduced participants' attention or affected their strategy use during this study (see further discussion below). Additional validation studies can compare face-to-face versus online environments in finer-grained detail.

Additionally, the challenges posed by learning to criterion might also incentivize online participants (who are unsupervised) to use different strategies than face-to-face participants do. Future validation studies can document and compare the strategies and behaviors of learners in online vs. face-to-face environments. For example, a study can ask participants to self-report their learning strategies, either retrospectively or concurrently. Notably, Bahrick and Hall (2005) asked participants to self-report the memorization strategy they used on a trial-by-trial basis. Data from either method might better link learning processes and learning outcomes.

Furthermore, as pointed out by an anonymous reviewer, the imbalanced recycling frequency of the dropout procedure in Training Phase 1 (words correctly answered twice were dropped from the pool) might have affected learning. As noted above, our experimental program maintained a maximum separation between items, but when few items remained in the pool (e.g., three items), the maximum separation was small (in this case, two items). Such small separation at later stages can potentially reduce retrieval difficulty and hinder long-term retention (Kornell & Bjork, 2008). Nevertheless, students studying vocabulary with flashcards often use such a dropout procedure (Wissman, Rawson, & Pyc, 2012) to spend more time learning the words that are difficult for them, so this procedure has ecological validity. When using such a methodology, however, participants' encoding of target words into their memory can vary substantially, which further highlights the need for data analysis at the item level.

Online research

Our study also examined the viability and validity of using an online platform for a multisession experimental study. Using pretests and benchmarking participant performance in our study against training session data (number of trials) from the original study, we excluded extremely high performing participants. We also excluded two participants who failed attention check items and eight participants (or bots) who failed bot checks (Newman et al., 2021).

Our study's attrition rate far exceeded the original study's attrition rate (61% > 15%), posing the greatest challenge to our online study's viability and validity. Among participants who dropped out of our study, 45% did so after Training Session 1 (after receiving the first payment). Possible causes of this high attrition rate include the difficulty of the first Training Session, substantial partial payment (£6) after the first

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

Training Session, or relatively low monetary incentive to return (another £6 after completing all three sessions).

Notably, longer or more difficult online experiments can increase attrition (see Rodd's [2024] review of potential threats to data quality in online experimental studies). As noted above, this study's participants required 166 trials (SD = 70) on average to learn 40 items to the criterion. Serfaty and Serrano (2024), in contrast, reported no attrition of participants in their multisession online study. With regard to difficulty, the training phase of their study also required participants to learn to criterion but only asked participants to learn 12 items to the criterion, which required a mean of only 19 trials (SD = 8). Regarding study length, both Serfaty and Serrano (2024) and the current study were similar with regard to posttest timing. Serfaty and Serrano based the timing of their posttest on Prolific's payment guidelines, which require participants to be compensated within 22 days of the end of the first experimental session. This allowed them to delay full payment until participants had completed the entire study. To retain fidelity to Cepeda et al.'s original study, follow Prolific payment guidelines, and operate within our budget constraints, we made two equal payments to participants. However, for longitudinal or multisession studies, where possible, Rodd recommends higher payments at later stages of the experiment to incentivize completion of the study. Overall, our study illustrates how the payment requirements of Prolific and other crowdsourcing programs may impact multiple aspects of an online study (e.g., participant motivation, attrition). In this sense, deciding how to structure payments or participant rewards is an important methodological consideration for multisession or longitudinal online studies.

The relatively high attrition rate in this study may also be due to the nature of online experimentation. Dropout rates can be dramatically higher for online studies than in-person experiments. Tomczak et al.'s (2023) meta-analysis of Gorilla's metadata showed an overall completion rate of 67.5%. Among single-session social psychology studies, attrition is much higher in online studies (up to 30%) than in equivalent in-person studies (typically 0%; Zhou & Fishbach, 2016). Attrition is also more likely to impact longitudinal or multisession online studies than one-shot experimental studies (Rodd, 2024).

Despite random assignment, high attrition can lead to *selective attrition* (or *attrition bias*; Rodd, 2024). This is a confound where attrition can lead to systematic differences between those who complete the study and those who do not. As noted, we address this potential confound by documenting attrition (Table 1), recording partial data (following suggestions from Rodd, 2024), and comparing the data and attributes of participants who dropped out of the study against those who completed it (see Validity section, above).

Although the final data set had high statistical power (.99 at the item level and .80 at the individual level), it was both costly (£4,230 [direct participant rewards + 25% Prolific overheard]) and required extensive time to screen the data, clean them, and exclude unacceptable data. To account for data quality issues in a one-time experiment, Uittenhove et al. (2023) suggest recruiting 20% additional participants for an online study than a face-to-face study. In contrast, our three-time-point study required 250% more participants (250% = [569 – 162]/162). As online multisession, longitudinal studies can have high attrition rates, researchers must plan accordingly.

Conclusion

The results of the mixed-effects model in this study showed evidence of spacing effects (advantage of a gap vs. no-gap condition) but no evidence of lag effects (advantages of longer gaps vs. shorter gaps), consistent with findings in the spacing literature (e.g., Kim

https://doi.org/10.1017/S0272263124000706 Published online by Cambridge University Press

& Webb, 2022). By way of conclusion, we would like to briefly comment on the pedagogical implications and generalizability of our study's findings.

Many researchers have raised examples of how the methodology typically employed in laboratory-based spacing research lacks ecological validity in relation to instructed L2 contexts (e.g., Marsden & Hawkes, 2024; Rogers & Cheung, 2020, 2021). The methodology of the current study required participants to study 40 word pairs at one time. This method arguably has some ecological validity because some popular flashcard apps for smartphones, such as *Quizlet*, *WordHolic*, *iKnow*, and *mikan*, allow learners to study 40 or more items at once. Studying a large number of items is also beneficial for learning because it introduces longer within-session spacing, which has been found to facilitate retention (Kim & Webb, 2022; Nakata, Suzuki, & He, 2023; Nakata & Webb, 2016).

However, with regard to classroom settings, many students (and teachers) would balk at trying to learn 40 unknown vocabulary items in a single lesson. Similar concerns have been raised by Marsden and Hawkes (2024) regarding the ecological validity of other aspects of spacing research, including whether the ISI–RI ratios generalize to scale with regard to school curricula, which necessitate the learning of real-world bodies of knowledge and cumulative final examinations covering material taught at different times (and intervals) throughout the academic year. Hence, future studies might consider such issues to increase the relevance and applicability of spacing research for L2 pedagogy.

To our knowledge, this is among the first multisession, experimental L2 studies to use crowdsourcing platforms (e.g., Serfaty & Serrano, 2024). The likely growth of experimental studies using online platforms and crowdsourcing data highlights the need for more methodologically-oriented research to validate the use of these platforms for L2 research. Such research can highlight the strengths and limitations of these platforms and scrutinize their generalizability and practicality across a range of experimental designs, research contexts, and population samples, thereby informing the design and quality of future such studies.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000706.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566–577. https://doi.org/10.1016/j.jml.2005.01.012

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *32*(2), 435–452. https://doi.org/10.1017/S0142716410000470

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354. https://doi.org/10.1037/0033-2909.132.3.354

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*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cepeda, N.J., Coburn, N., Rohrer, D., Wixted, J.T., Mozer, M.C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236–246. https://doi.org/10.1027/1618-3169.56.4.236

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2020). MLwiN version 3.05. Centre for Multilevel Modelling, University of Bristol.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, *40*(1), 83–107. https://doi.org/10.2307/40264512

Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in second language acquisition: Introduction to the special issue SLA for all? *Language Learning*, *73*(4), 981–1002. https://doi.org/10.1111/lang.12620

Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis*. Routledge.

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

Kim, K. M., Liu, X., Isbell, D. R., & Chen, X. (2024). A comparison of Lab- and Web-based elicited imitation: Insights from explicit-implicit L2 grammar knowledge and L2 proficiency. *Studies in Second Language Acquisition*, 1–22. doi:10.1017/S0272263124000214

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

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory*, *16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*(3), 373–388. https://doi.org/10.1007/s11251-013-9285-2

Laufer, B. (2012). Second language word difficulty. In C. A. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 5151–5156). Wiley-Blackwell.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27–44). Multilingual Matters.

Marsden, E., & Hawkes, R. (2024). Situating practice in a limited-exposure, foreign languages school curriculum. In Y. Suzuki (Ed.), *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology* (pp. 89–119). Routledge.

Mathôt, S., & March, J. (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning 72*(4), 1017–1048. https://doi.org/10.1111/lang.12509

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*, 916–939. https://doi.org/10.1017/S0272263121000292

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

Nakata, T., Suzuki, Y., & He, X. (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*, *73*(3), 799–834. https://doi.org/10.1111/lang.12553

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology*, *70*(3), 1380–1402. https://doi.org/10.1111/apps.12302

Patterson, A.S., & Nicklin, C. (2023). L2 self-paced data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*(1), 1–15 [advance online publication]. https://doi.org/10.1016/j.rmal.2023.100045

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113–138. https://doi.org/10.1177/1362168814568131

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Porte, G. K., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.
Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002
Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high-quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 1–20. https://doi.org/10.1016/j.jml.2023.104472
Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*(4), 857–866. http://www.jstor.org/stable/43893791
Rogers, J. (2021). Input spacing in second language classroom settings: Replications of Bird (2010) and Serrano (2011). *Language Teaching*, *54*(3), 424–433. https://doi.org/10.1017/s0261444820000439
Rogers, J. (2023). Spacing effects in task repetition research. *Language Learning*, *73*(2), 445–474. https://doi.org/10.1111/lang.12526
Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, *24*(5), 616–641. https://doi.org/10.1177/1362168818805251
Rogers, J., & Cheung, A. (2021). Does it matter when you review?: Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, *43*(5), 1138–1156. https://doi.org/10.1017/S0272263120000236
Ruiz, S., Chen, X., Rebuschat, P., & Meurers, D. (2019). Measuring individual differences in cognitive abilities in the lab and on the web. *PLoS One*, *14*, 226217. https://doi.org/10.1371/journal.pone.0226217
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
Serrano, R. (2022). A state-of-the-art review of distribution-of-practice effects on L2 learning. *Studies in Second Language Learning and Teaching*, *12*(3), 355–379. https://doi.org/10.14746/ssllt.2022.12.3.2
Serrano, R., & 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*(4), 971–994. https://doi.org/10.1002/tesq.445
Serrano, R., & Huang, H. Y. (2023). Time distribution and intentional vocabulary learning through repeated reading: a partial replication and extension. *Language Awareness*, *32*(1), 1–18. https://doi.org/10.1080/09658416.2021.1894162
Serfaty, J., & Serrano, R. (2024). Practice makes perfect, but how much is necessary? The role of relearning in second language grammar acquisition. *Language Learning*, *74*(1), 218–248. https://doi.org/10.1111/lang.12585
Suzuki, Y., & DeKeyser, R. (2017). Exploratory research on second language practice distribution: An Aptitude× Treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56. https://doi.org/10.1017/S0142716416000084
Tomczak, J., Gordon, A., Adams, J., Pickering, J. S., Hodges, N., & Evershed, J. K. (2023). What over 1,000,000 participants tell us about online research protocols. *Frontiers in Human Neuroscience*, *17*, 1228365. https://doi.org/10.3389/fnhum.2023.1228365
Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, *6*(1), 1–13. https://doi.org/10.5334/joc.259
Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L.J., & Verschere, B. (2024). The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. *Behavior Research Methods*, *56*, 4836–4849. https://doi.org/10.3758/s13428-023-02221-2
Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from readinga graded reader? *Reading in a Foreign Language*, *15*, 130–163. http://doi.org/10.1177/003368828501600214
Wiseheart, M., Kim, A. S. N., Kapler, I. V., Foot-Seymour, V., & Kupper-Tetzel, C. E. (2019). Enhancing the quality of student learning using distributed practice. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 550–584). Cambridge University Press.
Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579. https://doi.org/10.1080/09658211.2012.687052
Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Cite this article: Rogers, J., Nakata, T., & Chiu, M. M. (2025). Optimizing distributed practice online: A conceptual replication of Cepeda et al. (2009). *Studies in Second Language Acquisition,* 47: 417–439. https://doi.org/10.1017/S0272263124000706