Mem Cogn
DOI 10.3758/s13421-014-0406-1

# Between-list lag effects in recall depend on retention interval

**Mary A. Pyc · David A. Balota · Kathleen B. McDermott · Tim Tully · Henry L. Roediger III**



**Abstract** Although the benefits of spaced retrieval for long-term retention are well established, the majority of this work has involved spacing over relatively short intervals (on the order of seconds or minutes). In the present experiments, we evaluated the effectiveness of spaced retrieval across relatively short intervals (within a single session), as compared to longer intervals (between sessions spaced a day apart), for long-term retention (i.e., one day or one week). Across a series of seven experiments, participants ($N$ = 536) learned paired associates to a criterion of 70 % accuracy and then received one test–feedback trial for each item. The test–feedback trial occurred within 10 min of reaching criterion (short lag) or one day later (long lag). Then, a final test occurred one day (Exps. 1–3) or one week (Exps. 4 and 5) after the test–feedback trial. Across the different materials and methods in Experiments 1–3, we found little benefit for the long-lag relative to the short-lag schedule in final recall performance—that is, no lag effect—but large effects on the retention of information from the test–feedback to the final test phase. The results from the experiments with the one-week retention interval (Exps. 4 and 5) indicated a benefit of the long-lag schedule on final recall performance (a lag effect), as well as on retention. This research shows that even when the benefits of lag are eliminated at a (relatively long) one-day retention interval, the lag effect reemerges after a one-week retention interval. The results are interpreted within an extension of the bifurcation model to the spacing effect.

**Keywords** Memory · Recall · Spacing effects · Lag effects

M. A. Pyc (✉) · D. A. Balota · K. B. McDermott · H. L. Roediger III
Department of Psychology, Washington University, One Brookings Drive, Box 1125, St. Louis, MO 63130, USA
e-mail: mpyc@wustl.edu

T. Tully
Dart Neuroscience, San Diego, CA, USA

Psychologists have studied dozens of variables that affect learning and retention, but the distribution or spacing of practice is one of the most venerable and well documented. Seminal work by Ebbinghaus (1885/1913) demonstrated the powerful effects of spacing learning events, showing enhanced performance when practice trials are spaced across time and/or other materials (i.e., spaced practice) as compared to when they occur consecutively (i.e., massed practice). In subsequent years, hundreds of studies have documented the memorial benefits of spacing (commonly referred to as the spacing effect; for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Crowder, 1976; Donovan & Radosevich, 1999; Greene, 2008). Spacing effects have been documented across the lifespan (e.g., Balota, Duchek, & Paullin, 1989; Kornell, Castel, Eich, & Bjork, 2010; Sobel, Cepeda, & Kapler, 2011; Toppino, Fearnow-Kenney, Kiepert, & Teremula, 2009), across species (e.g., Robbins & Bush, 1973; Tully, Preat, Boyton, & Del Vecchio, 1994), and in patient populations (e.g., Balota et al., 2006; Goverover, Basso, Wood, Chiaravalloti, & DeLuca, 2011).

The most common spacing effect paradigm in the tradition of research with humans involves presenting participants a list of to-be-learned items—words, pictures or other events—some of which are repeated and others not. In one standard comparison, researchers include two types of repeated items: those massed and those distributed or spaced (e.g., Madigan, 1969; Melton, 1967, 1970). In a massed presentation condition, the second study event occurs immediately after the first presentation, whereas in a spaced presentation condition the second study event occurs after some intervening amount of time and/or intervening number of items (e.g., after five other items). The general finding is that performance on a later retention test is higher for items that are spaced relative to those that are massed (the spacing effect). Within a spaced condition, a further manipulation is to vary the lag or spacing of items (e.g., 1, 3, 5, 10, or 20 items between repetitions). When lag affects performance, this is referred to as the lag effect (i.e., over and above any effect of spacing).

Published online: 19 March 2014

In addition to spacing, the effect of lag can be quite powerful. For example, Madigan (1969) not only found better free recall for spaced than massed items, but performance increased as the lag between the repetition of items increased, with about 15 % improvement in free recall between lags of 2 and 40. The lag effect is intriguing because it suggests that the benefits from spacing are not due to idiosyncratic aspects of the massed condition—that is, the word is repeated—and so may just receive less processing on the second occurrence. However, such deficient processing may represent part of the story; Madigan found little improvement between a single presentation and back-to-back presentations (massed practice), but adding a lag of just two items boosted recall about 10 % above the massed presentations.

Traditionally in studies of the lag effect, practice involves merely studying items, but robust lag effects have also been shown when practice involves studying and then testing of items at various lags (with or without feedback; e.g., Cull, 2000; Karpicke & Bauernschmidt, 2011). In this case, items benefit not only from lag but also from retrieval practice (for recent reviews on retrieval practice effects, see Rawson & Dunlosky, 2011, and Roediger & Butler, 2011).

Although the lag effect is robust, the vast majority of previous research has involved lags that are relatively short, with only seconds or minutes between repetition trials, and in experiments that usually do not involve learning to a criterion during study (but see Pyc & Rawson, 2009). The present experiments were designed to extend this research to examine how relatively long lag intervals (e.g., 10 min vs. 24 h) influence retention. Surprisingly, relatively little research has evaluated this question (for examples of the few exceptions to this, see Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Glenberg & Lehmann, 1980; Litman & Davachi, 2008; Küpper-Tetzel & Erdfelder, 2012; Simone, Bell, & Cepeda, 2012).

In the present experiments, we were interested in examining learning schedules that students might naturally use during self-regulated learning. Imagine a student studying for an upcoming exam. Although students can study in many ways (e.g., rereading the text or studying in a group; e.g., Karpicke, Butler, & Roediger, 2009), many report using flashcards on a regular basis (Hartwig & Dunlosky, 2012; Wissman, Rawson, & Pyc, 2012). Using flashcards to study allows students to use an effective learning strategy: self-testing with feedback. In studying critical concepts, students may write a term on one side of the flashcard and the definition on the back. During practice, a student might first try to retrieve the correct answer given the term and then flip the card over to determine whether their answer was correct. Many studies have shown that testing followed by feedback is an effective study strategy for improving long-term retention (see Rawson & Dunlosky, 2011, and Roediger & Butler, 2011), and students report using self-testing to determine how well they know information (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007). Thus, in the present study, instead of repeated studying, as in most studies of lag effects, we used a test–feedback procedure in which participants attempted to retrieve the correct response to a cue and then received the answer as feedback.

In summary, previous research has established that (1) increasing the lag between repetitions is beneficial for long-term retention, (2) testing followed by feedback (hereafter referred to as *test–feedback practice*) is beneficial for long-term retention, and (3) students sometimes use test–feedback practice (e.g., via flashcards) during self-regulated study. What is less well understood from the extant literature is the extent to which the interval (lag) between test–feedback episodes matters for long-term retention. Most of the literature has used short lags (e.g., Landauer & Bjork, 1978). If the lag between initial study and test–restudy is on the order of days as opposed to minutes, will long-term retention be greatly boosted? The present experiments were designed to evaluate the effectiveness of test–feedback learning episodes separated by various lags, to answer this question. We had participants learn information to a criterion, followed by one extra trial of test–feedback practice. The final trial of practice occurred either about 10 min or about one day after learning to criterion. Then the final criterial test occurred either one day (Exps. 1–3 and 5) or one week (Exps. 4 and 5) after the final test–feedback practice trial. To presage the results, we obtained no lag effect at the one-day retention interval, but we did at one week.

## General method

The method across each of the experiments was largely the same, so we provide an overview of the general method here (see Fig. 1). Any variation will be described in the Method section of the relevant experiment. Participants were recruited via Amazon.com's Mechanical Turk (mTurk), an online crowd-sourcing platform that has become increasingly popular among researchers (see Mason & Suri, 2012, for an evaluation of mTurk and a description of how to use it for research purposes; see also Sargis, Skitka, & McKeever, 2013). People on mTurk complete tasks and are awarded credit toward earning a gift card for successfully completing tasks (researchers pay Amazon for each participant, and the participant is awarded that amount toward a gift card). All results reported below are based on participants who reported being residents of the United States with English as their primary language. They had a quality rating of 80 % or greater, indicating the percentage of time that their work was approved once submitted (requesters have the ability to accept or reject a person's work—e.g., a participant might not complete the full task). Participants were paid $10 for fully completing the present experiments. We discuss exclusion

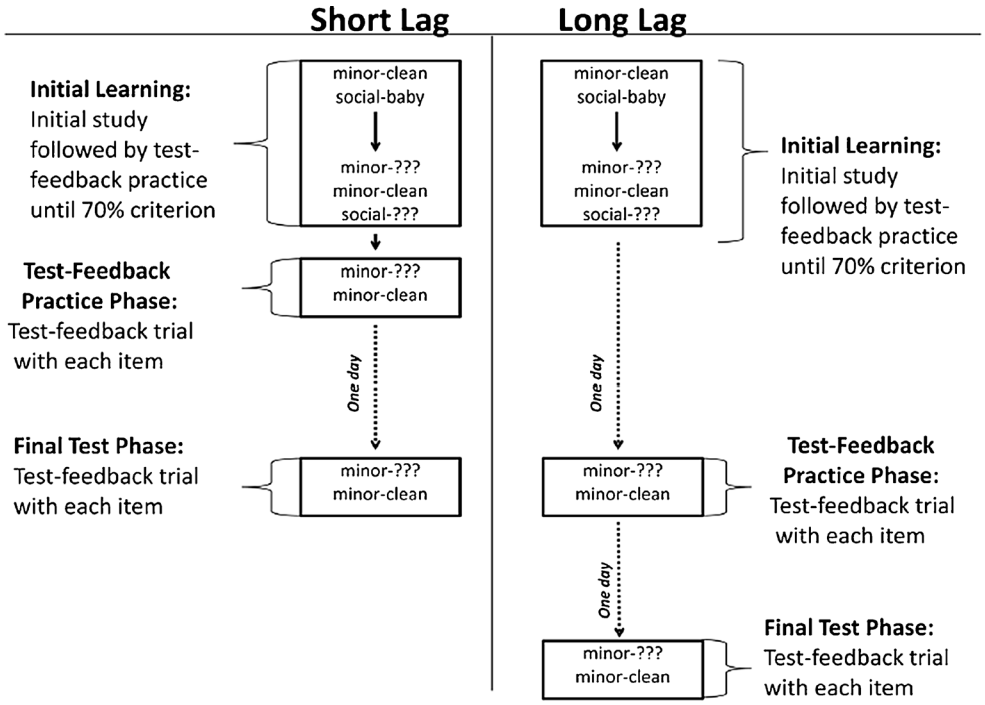


**Fig. 1** General method of all experiments

criteria for the experiments below, based on participant attrition and related issues.

Participants learned pairs of items across the three phases in a standard paired-associate learning paradigm. The three phases were the initial learning phase, the test–feedback phase—which constituted a test with an immediate restudy opportunity for the pair just after the test—and the final test phase. During the initial learning phase, all pairs (A–B) were presented individually on a computer screen for an initial study trial. After the initial study, participants engaged in cycles of tests with feedback, with all pairs being tested on each trial. This procedure continued until 70 % of the to-be-learned items were correctly recalled on a given cycle during acquisition (hereafter referred to as the *criterion cycle*). For test trials, the cue was presented (A) and participants were asked to retrieve the associated response (B) and to type it on the keyboard. Feedback immediately followed the test trial for a given item (i.e., the cue and target were restudied), regardless of the retrieval success on the preceding test trial. If fewer than 70 % of the items in the list were recalled, participants received an additional cycle of test–feedback practice until the 70 % criterion was achieved.

After reaching criterion, participants moved on to the next part of the experiment: the test–feedback phase. During this phase, each item was presented for one test trial and immediately followed by feedback. Critically, the test–feedback phase occurred either after a short lag (immediately after the criterion phase in some experiments, or after 10 min in Exp. 3) or a long lag (24 h later, in a different session). Lag was manipulated between subjects, with random assignment of participants to conditions. Finally, for both groups, the final test phase occurred one day after the test–feedback phase (i.e., Day 2 for the short-lag group, and Day 3 for the long-lag group). Thus, the spacing of repetitions occurred either immediately after the initial learning session or after 24 h, and the retention interval was 24 h in Experiments 1–3. The general procedure in Experiments 4 and 5 was the same, with the exception that the retention interval between the last test–feedback cycle and the final test was varied. Experiment 4 involved a one-week retention interval, and Experiment 5 involved 20-min, one-day, or one-week intervals. The final (criterial) test always involved a single test trial for each item, which was immediately followed by feedback (we presented feedback on the final test even though participants were not tested again).

All participants were required to complete three sessions, even though only the first two sessions were of interest in the short-lag group. This requirement was made to ensure that all participants had to meet similar requirements to successfully complete the study, and therefore that participants could not self-select on the basis of the demands of the longer spacing and/or retention interval conditions. For the short-lag group, the third session occurred one day after the criterial test in the second session, and again consisted of one test–feedback trial for each item. We do not report these data because they are not of interest for present purposes.

We opted to equate the retention intervals between the end of learning (i.e., the test–feedback practice phase) and the final test phase so that we could evaluate the extent to which forgetting differed between the lag groups across equated retention intervals. Of course, we could have equated the total

time from the beginning of learning (during the initial learning phase) to the final test phase, but lag would then have been confounded with retention interval. That is, a shorter amount of time would elapse between the test–feedback trial and the final test phase for the long-lag versus the short-lag group. If the long-lag group were to perform better than the short-lag group, we could not evaluate whether this outcome was due to the lag schedule or to the differing retention intervals. For this reason, we opted to equate the retention intervals across conditions.

All participants completed three sessions (even though only data from the first two sessions were used for the short-lag group). After the final criterial test, participants reported whether they had written down any words during any of the prior sessions. The question was phrased with the intention that participants would not believe that they had done something wrong if they wrote anything down (i.e., we had not specifically told them not to write anything down, but of course we hoped they would not). Specifically, we asked "During any of the sessions, did you happen to note down any of the items you were asked to learn? If so, how many items did you write down?" Although we cannot be sure that all participants accurately reported the extent to which they wrote down items, we feel fairly confident that their reporting was accurate, because 8 % of participants reported such activity, and they often provided an explanation for why they wrote things down. Across all experiments, 80 % of the individuals who initially began the tasks provided usable data by meeting our two criteria of completing all three sessions and reporting that they had not written anything during the sessions. We lost 12 % of participants, who failed to complete all three sessions, as well as those who reported writing material (the 8 % reported above). The attrition rates were nearly the same for the short- and long-lag conditions (collapsed across the experiments, 37 short-lag and 39 long-lag individuals failed to complete all three sessions).

## Experiment 1

### Method

*Participants and design* A group of 82 people participated, and were randomly assigned to the short-lag ($N = 42$) or the long-lag ($N = 40$) groups.

*Materials* The items included 35 unrelated word pairs.

*Procedure* The overall procedure was the same as was described in the General Method. During the initial study trial, the cue and target were presented together for 5 s (e.g., beach–risk). For test trials, only the cue was presented (beach–????), and participants had 6 s to retrieve and type the target answer (risk). Immediately after 6 s had elapsed, the cue and target were presented together as feedback for 3 s.

### Results and discussion

For each experiment, we report the results in two sections. First, we report two measures of initial learning performane:, the mean number of trials to reach the 70 % learning criterion, and the percentage of recall on the last test trial during this initial learning phase. These measures permit an examination of whether group differences existed prior to the lag manipulation (they should not, of course). We then report performance on the tests during the test–feedback phase (same day or next day), and finally report recall on the final (criterial) test, the measure of primary interest.

*Initial learning performance* As expected, we found no difference for the short- and long-lag groups in trials to criterion during the initial learning phase [2.98 vs. 3.45, respectively; $F(1, 81) = 1.91$, $p = .17$, $\eta^2 = .02$], nor for performance on the criterion trial (.78 vs. .78, $F < 1$).

*Performance during the test–feedback and final test phases* Performance on the test trial during the test–feedback phase is presented in Table 1, where it can be seen that performance was much better immediately after learning to criterion (.90; short lag) than a day later (.61; long lag), $F(1, 80) = 105.47$, $p < .001$, $\eta^2 = .57$. This outcome is hardly a surprise—forgetting occurs over a day. Our primary interest was how the test and the feedback after the test would affect recall a day later. The vast literature on the lag effect might lead one to expect that final performance would be much better in the long-lag than in the short-lag condition, but we did not obtain this outcome. As can be seen in Fig. 2 (and Table 1), little difference was apparent between the two conditions after a 24-h retention interval, $F < 1$. This outcome is surprising, given that the benefits of repetition at long lags relative to short lags are typically quite robust with long retention intervals. Lag almost always has an effect, even with much shorter lags than 24 h. Hence, this exception to the standard outcome deserves careful scrutiny, which we have provided in additional experiments.

Performance dropped over the 24-h retention interval in the short-lag group, but no drop in performance occurred in the long-lag group (see Table 1). To evaluate the extent to which the lag manipulation differentially influenced recall across the one-day retention interval, we conducted a 2 (lag: short vs. long) × 2 (test type: test–feedback trial vs. final test trial) mixed-factor analysis of variance (ANOVA). The main effects of test type and lag were significant, $F(1, 80) = 14.17$, $p < .001$, $\eta^2 = .09$, and $F(1, 80) = 14.45$, $p < .001$, $\eta^2 = .15$, respectively. The interaction was also significant [$F(1, 80) = 66.81$, $p < .001$,

**Table 1** Mean proportions of items recalled on test trials during test–feedback and final test phases

| | Test–Feedback | Final Test | Difference Score* |
|---|---|---|---|
| Experiment 1 | | | |
| Short lag | .90 (.01) | .66 (.03) | –.24 |
| Long lag | .61 (.03) | .70 (.03) | .09 |
| Experiment 2a | | | |
| Short lag | .93 (.02) | .83 (.03) | –.10 |
| Long lag | .70 (.05) | .84 (.04) | .14 |
| Experiment 2b | | | |
| Short lag | .93 (.01) | .91 (.02) | –.02 |
| Long lag | .84 (.02) | .91 (.02) | .07 |
| Experiment 2c | | | |
| Short lag | .91 (.02) | .83 (.04) | –.08 |
| Long lag | .65 (.05) | .77 (.05) | .12 |
| Experiment 3 | | | |
| Short lag | .81 (.03) | .79 (.05) | –.02 |
| Long lag | .67 (.03) | .84 (.03) | .17 |
| Experiment 4 | | | |
| Short lag | .93 (.02) | .63 (.04) | –.30 |
| Long lag | .86 (.02) | .75 (.04) | –.11 |
| Experiment 5 | | | |
| Short lag (20-min RI) | .75 (.03) | .80 (.03) | .05 |
| Long lag (20-min RI) | .55 (.05) | .81 (.04) | .26 |
| Short lag (1-day RI) | .81 (.03) | .70 (.04) | –.11 |
| Long lag (1-day RI) | .55 (.04) | .71 (.04) | .16 |
| Short lag (1-week RI) | .77 (.03) | .28 (.05) | –.49 |
| Long lag (1-week RI) | .50 (.04) | .42 (.04) | –.08 |

Values in parentheses represent standard errors. * Negative values indicate a decrease in performance from the test–feedback to the final test trials, whereas positive values indicate an improvement in performance from the test–feedback to the final test trials. Readers should be cautious in interpreting this measure, given that participants were at different levels of performance during the test–feedback practice phase and we do not have a measure of performance after the feedback portion of this phase

$\eta^2$ = .42], with greater differences in recall between the test–feedback phase and the final test phase for the short-lag relative to the long-lag group. That is, forgetting occurred across the 24-h retention interval when the lag between learning episodes occurred within one session [i.e., the short-lag group, $t(41) = 6.83$, $p < .001$, $d = 2.13$], but improvement occurred when the lag occurred after an interval of a day [i.e., long-lag group, $t(39) = 4.85$, $p < .001$, $d = 1.55$]. However, these data do not represent the usual forgetting curve between two retention intervals, because participants received feedback after the first test. Thus, another way to view the results is that the feedback prevented forgetting in the long-lag but not in the short-lag condition. Of course, performance was also much higher during the test–feedback phase in the short-lag than in the long-lag condition, but the important point is that the directions of the difference were opposite in the two cases.

## Experiments 2a, 2b, and 2c

Experiment 1 demonstrated that the lag between the initial learning phase and the test–feedback phase does not enhance performance on a delayed final test (i.e., no lag effect occurs), but that the lag does influence retention between the test–feedback phase and final test phase. The absence of a lag effect was unexpected given how well-documented such effects have been from the earliest days of memory research. Because the paradigm used in Experiment 1 differs from most previous research on lag effects, we examined whether the null effect would replicate using different materials (i.e., face–name paired associates) with the same general method used in Experiment 1. Experiment 2a, 2b, and 2c are presented together here because the only difference between each was the number and specific combination of face–name pairs that were presented to participants. Across

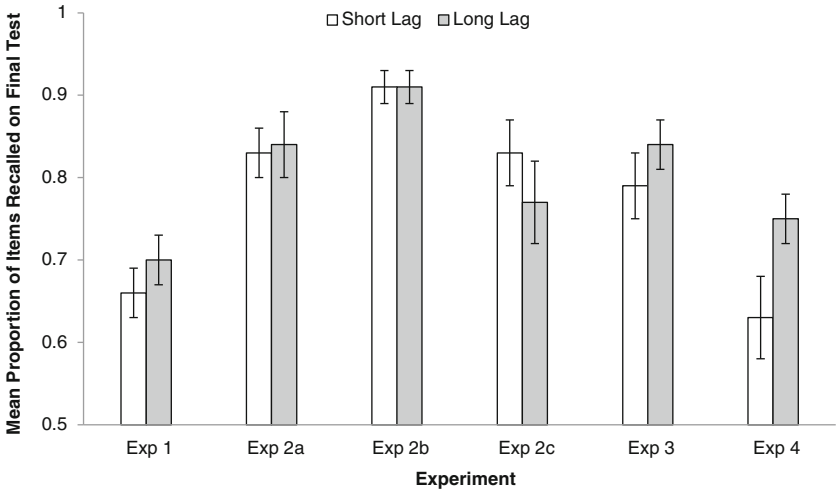


**Fig. 2** Mean proportions of items recalled on final test trials for Experiments 1, 2, 3, 4. Error bars represent standard errors

experiments, the number of to-be-learned pairs was increased from ten to 24 from Experiment 2a to 2b. Experiment 2c required participants to learn 24 face–name pairs of people of only one gender in order to make the task harder yet.

## Method

*Participants and design* Across the three experiments, 153 individuals were recruited to participate (Exp. 2a, short-space $N = 29$, long-space $N = 27$; Exp. 2b, short-space $N = 24$, long-space $N = 30$; Exp. 2c, short-space $N = 27$, long-space $N = 16$). Participants were randomly assigned to the short-lag group or the long-lag group.

*Materials* In Experiment 2a, the items included ten face–name pairs (half male and half female); Experiment 2b included 24 face–name pairs (half male and half female); and Experiment 2c included 24 face–name pairs (all male). All faces were downloaded from the Psychological Image Collection at Stirling (http://pics.psych.stir.ac.uk). First and last names were generated by searching through a local phone book for first and last names of medium frequency and then combining the first and last names (e.g., Brandon Irwin) so that common or unique names were avoided. These items were taken from Maddox and Balota (2012).

*Procedure* The procedures for Experiments 2a, 2b, and 2c were the same. For the initial study trial, each face–name pair was presented individually on the computer screen for 10 s. For test trials, participants received 15 s to retrieve the appropriate name when prompted with the cue (i.e., the face). For Experiment 2a, participants were asked to learn the first and last names associated with a face, whereas for Experiments 2b and 2c, participants were only asked to learn first names. Immediately after 15 s had elapsed on test trials, participants received a 6-s restudy trial with both the face and name. If participants retrieved the name before 15 s had elapsed, they could press a button to advance to the feedback trial for the pair. More time was allotted for study and test trials for face–name pairs (relative to the word pairs used in Exp. 1) because, on the basis of other research in our lab, they are generally more difficult to learn, and participants are slower to retrieve the names when given the face than to retrieve word pairs.

## Results and discussion

*Initial learning performance* We again found no differences between the two different lag groups for the initial learning phase. The mean numbers of cycles to criterion were similar for the short- and long-lag groups for Experiment 2a (2.38 vs. 2.26, $F < 1$), Experiment 2b (2.04 vs. 1.90, $F < 1$), and Experiment 2c [2.52 vs. 3.06; $F(1, 41) = 1.25$, $p = .270$]. Similarly, performance on the criterion trial did not differ between the short- and long-lag groups in Experiment 2a (.84 vs. .83, $F < 1$), Experiment 2b (.84 vs. .86, $F < 1$), or Experiment 2c (.81 vs. .79, $F < 1$).

*Performance during the test–feedback and final test phases* Performance on the test trial during the test–feedback phase is presented in Table 1 for Experiments 2a, 2b, and 2c. The same pattern of forgetting occurred for the long relative to the short lag in each case. For each experiment, we conducted a between-subjects ANOVA, which yielded significant differences between the short- and long-lag groups [Exp. 2a, $F(1, 54) = 18.91$, $p < .001$, $\eta^2 = .26$; Exp. 2b, $F(1, 52) = 12.06$, $p = .001$, $\eta^2 = .19$; Exp. 2c, $F(1, 41) = 32.69$, $p < .001$, $\eta^2 = .44$].

More importantly, we obtained no lag effect in any of the experiments, as can be seen in Fig. 2. Between-subjects ANOVAs indicated no significant difference for proportions recalled on the final test as a function of lag in any of the experiments, all $F$s < 1.

Replicating Experiment 1, we again found that performance dropped over the 24-h retention interval in the short-lag groups, but not in the long-lag groups (see Table 1). For each experiment, we conducted a 2 (lag: short vs. long) × 2 (test type: test–feedback trial vs. final test trial) mixed-factor ANOVA, and the results were similar across experiments (and similar to those of Exp. 1). Specifically, a main effect of lag was obtained [Exp. 2a, $F(1, 54) = 5.74$, $p = .02$, $\eta^2 = .10$; Exp. 2b, $F(1, 52) = 4.41$, $p = .04$, $\eta^2 = .08$; and Exp. 2c, $F(1, 41) = 10.45$, $p = .002$, $\eta^2 = .20$]. The main effect of test type was not significant in Experiment 2a or 2c ($F$s < 1), but it approached significance in Experiment 2b, $F(1, 52) = 2.93$, $p = .09$, $\eta^2 = .04$. Most importantly, replicating the pattern of results found in Experiment 1, the interaction of lag and test type was highly significant in each experiment [Exp. 2a, $F(1, 54) = 20.56$, $p < .001$, $\eta^2 = .28$; Exp. 2b, $F(1, 52) = 13.24$, $p < .001$, $\eta^2 = .19$; Exp. 2c, $F(1, 41) = 22.34$, $p < .001$, $\eta^2 = .35$], indicating different retention functions for the short- and long-lag groups. For the short-lag group, we again found a decrease in performance between the test–feedback phase and the final test phase [Exp. 2a, $t(28) = 3.32$, $p = .003$, $d = 1.25$; Exp. 2b, $t(23) = 1.72$, $p = .10$, $d = 0.72$; Exp. 2c, $t(26) = 3.18$, $p = .004$, $d = 1.25$ (the decrease was not significant in Exp. 2b, but performance was near ceiling)]. For long-lag groups, performance again increased significantly [Exp. 2a, $t(26) = 3.11$, $p = .004$, $d = 1.22$; Exp. 2b, $t(29) = 3.46$, $p = .002$, $d = 1.29$; Exp. 2c, $t(15) = 3.67$, $p = .002$, $d = 1.90$].

# Experiment 3

Results from Experiment 2a–2c replicated and extended those of Experiment 1. Using face–name pairs, we found no difference in performance on the final test for short- and long-lag

groups. However, we did find that the long-lag group achieved better recall on the final test (relative to the test 24 h previously) than did the short-lag group (which again showed forgetting). That is, in each case performance from the test–feedback phase to the final test phase decreased for the short-lag group but increased for the long lag group. However, despite this consistent pattern, recall on the final test was never greater in the long-lag than in the short-lag condition—that is, no lag effect emerged. The consistent lack of a lag effect in these experiments is surprising, given the powerful effects of lag reported in the literature. Nonetheless, the absence of a lag effect in our conditions appears quite robust, replicating across four experiments using word-word paired associates or face–name pairs. Therefore, the final three experiments were conducted to further explore aspects of the design that may have led to this pattern. At the very least, we have found a boundary condition for lag effects.

Experiment 3 was designed to evaluate the possibility that the trial during the test–feedback phase was functionally part of the initial learning phase for the short-lag group (because it occurred immediately after criterion had been reached), whereas it was a distinct event for the long-lag group (because it occurred one day later). Hence, in addition to lag, the event structure of the experiment may have changed (see Zacks & Swallow, 2007). Thus, Experiment 3 was designed to increase the likelihood that all participants would experience the initial learning phase and the test–feedback phase as separate events (see Zacks, Speer, Swallow, Braver, & Reynolds, 2007). To do so, Experiment 3 included a 10-min delay between the initial learning phase and the test–feedback phase for the short-lag group (leading participants to perceive these phases as separate events), but maintained the 24-h interval for the long-lag group. Both short- and long-lag groups were then given a final test after 24 h, as in the previous experiments.

## Method

*Participants and design* A group of 63 individuals were randomly assigned to the short-lag group ($N = 28$) or the long-lag group ($N = 35$).

*Materials* Items included the ten face–name pairs used in Experiment 2a. We opted for only ten face–name pairs, in order to reduce the length of the first session because the patterns of results were similar, regardless of the number and composition of face–name pairs (as shown in Exps. 2a–2c).

*Procedure* The procedure was identical to that of Experiment 1, with the exception of the following changes. First, after items were learned to criterion during the initial learning phase, participants in both groups completed 10 min of trivia questions. For the trivia task, a question was presented (e.g., which poisonous substance is also known as "Wooly Rock"?), and participants received 10 s to type the correct answer (e.g., asbestos). No feedback was provided for trivia questions, and performance on these questions is not reported below. Second, after 10 min had elapsed, the short-lag group completed the test–feedback phase, whereas the long-lag group completed the test–feedback phase one day later. Thus, the trivia task served to increase the likelihood that the short-lag group experienced the initial learning phase and the test–feedback phase as distinct events.

## Results and discussion

*Initial learning performance* The mean numbers of cycles to criterion were similar for the short- and long-lag groups [2.64 vs. 3.14; $F(1, 61) = 1.71$, $p = .196$, $\eta^2 = .03$]. Similarly, performance on the criterion trial did not differ for the short- and long-lag groups [.84 vs. .79; $F(1, 61) = 2.03$, $p = .159$, $\eta^2 = .03$].

*Performance during the test–feedback and final test phases* Performance on the test trial during the test–feedback phase is presented in Table 1, where it is shown that forgetting occurred between 10 min and one day, $F(1, 60) = 8.87$, $p = .004$, $\eta^2 = .13$.

Performance on the final test for each group is presented in Fig. 2. As in all prior studies, performance on the final test did not differ between the short- and long-lag groups, $F < 1$. Thus, with a one-day retention interval, the nearly 24-h spacing between the initial learning phase and the test–feedback phase did not appear to differentially influence final test performance.

To evaluate the extent to which the lag manipulation differentially influenced retention between the test–feedback trial and the final test trial, we again conducted a 2 (lag: short vs. long) × 2 (test type: test–feedback trial vs. final test trial) mixed-factor ANOVA (see Table 1). The main effect of test type was significant [$F(1, 60) = 11.03$, $p = .03$, $\eta^2 = .12$], but the main effect of lag was not ($F < 1$). Importantly, the interaction was again significant [$F(1, 60) = 20.39$, $p < .001$, $\eta^2 = .22$].

Although the critical interaction remained, the introduction of a 10-min interval between the initial learning phase and the test–feedback phase for the short-lag group did influence the pattern of results for this condition. In all of the previous experiments (except Exp. 2b with its ceiling effect), performance decreased from the test–feedback phase to the final test phase for the short-lag group (all $p$s $< .001$), whereas in the present study, performance did not change between the test–feedback phase and the final test phase in the short-lag group, (a 2 % effect, $t < 1$). These results suggest that even a brief, 10-min interval (along with a change in task context) between encoding phases reduces forgetting across a delayed retention interval. This is probably because the 10-min interval before

the last test–feedback phase in Experiment 3 had already introduced some forgetting (the immediate test in Exp. 2a produced .93 recall, whereas the delay in Exp. 3 led to .81 recall). The test–feedback practice after some forgetting seems to have prevented further forgetting after a day. For the long-lag group, we again found a significant improvement in performance between the test–feedback and final test phases, $t(34) = 5.96$, $p < .001$, $d = 2.04$, replicating prior experiments.

## Experiment 4

Experiment 3 provided an important extension beyond the previous studies by replicating the critical interaction between lag and test type with a longer interval between the initial learning phase and the test–feedback phase for the short-lag group. This interval increased the likelihood that the learning phases would be experienced as distinct events (Zacks et al., 2007). Replicating results from the previous studies, when learning was spaced across a one-day interval (the long-lag schedule), recall improved from the test–feedback phase to the final test phase. However, when learning occurred within one session (the short-lag schedule), recall decreased from the test–feedback phase to the final test phase. In Experiments 1, 2a, and c this decrease was significant. In Experiment 3, with a 10-min interval between initial learning and test–feedback practice, this decrease in recall was no longer significant, suggesting that a 10-min interval in a test–restudy spaced condition can buffer against forgetting across a one-day retention interval.

Taken together, our results suggest that the short- and long-lag schedules have different forgetting functions, even though no difference occurs on the final test after one day. Assuming that the forgetting functions of the lag schedules do differ, then extending the retention interval should reveal differences in recall on the final test at long intervals. That is, a lag effect should emerge with a longer retention interval. To evaluate this hypothesis, the retention interval between the test–feedback phase and the final test phase was extended to one week. We predicted that this longer retention interval would finally produce a lag effect.

### Method

*Participants and design* A group of 54 individuals were randomly assigned to either the short-lag group ($N = 25$) or the long-lag group ($N = 29$).

*Materials* Items included the 24 face–name pairs (half male and half female) from Experiment 2b.

*Procedure* The procedure was identical to that of Experiment 1, with the exception that the retention interval between the test–feedback phase and the final test phase was one week.

### Results and discussion

*Initial learning performance* The mean numbers of cycles to criterion were similar for the short- and long-lag groups (2.12 vs. 1.90, $F < 1$). Similarly, performance on the criterion trial did not differ for the short- and long-lag groups (.84 vs. .85, $F < 1$).

*Performance during the test–feedback and final test phases* Performance on the test trial during the test–feedback phase is presented in Table 1 and shows greater forgetting on the test–feedback trial for the long- relative to the short-lag group, $F(1, 52) = 5.44$, $p = .02$, $\eta^2 = .10$. The data are nearly the same as in Experiment 2b, making the results between experiments quite comparable.

Of primary interest, we obtained a significant lag effect in Experiment 4 (see Fig. 2), with performance on the final test being greater for the long- than for the short-lag group, $F(1, 52) = 5.02$, $p = .03$, $\eta^2 = .09$. This is the first time that we have shown a lag effect in this series of experiments, so the retention interval after the test–feedback phase seems to be the critical variable.

Table 1 shows performance on the test–feedback trial and final test trial for both the short- and long-lag groups, and for the first time we observe forgetting for both the short-lag and long-lag conditions ($t$s > 4.3). Of course, this is not surprising, since the retention interval increased to one week as opposed to one day in the previous studies. The results of a 2 (lag: short vs. long) × 2 (test type: test–feedback trial vs. final test trial) ANOVA showed a main effect of test type, $F(1, 52) = 85.62$, $p < .001$, $\eta^2 = .55$, but the main effect of lag was not significant, $F < 1$. Importantly, the interaction was again significant, $F(1, 52) = 16.98$, $p < .001$, $\eta^2 = .11$. Although performance was lower overall after a one-week retention interval, we found significantly more forgetting from the test–feedback phase to the final test phase in the short-lag than in the long-lag group.

## Experiment 5

Experiments 1–3 yielded different retention functions for the short- and long-lag groups, but performance was consistently similar on the final tests for the two groups. By extending the retention interval to one week in Experiment 4, we were able to show that differences in retention function lead to differences in final test performance when the retention interval is sufficiently long (i.e., one week instead of one day). However, a skeptic might wonder whether this last result would be replicable and whether retention interval was the critical variable. The goal of this final experiment was to replicate and extend the results from Experiment 4 by including additional retention interval groups, so as to show our pattern of results from across experiments within a single experiment. Thus, in

addition to the one-day and one-week groups of the prior studies, Experiment 5 also included a 20-min retention interval group. One might expect the lag effect to reverse at this short interval (see Glenberg & Lehmann, 1980). Furthermore, to ensure that the pattern of results that we have observed across all of the prior studies was not due to the particular criterion level that items were learned to (i.e., 70 %), Experiment 5 required that items be learned to a 50 % criterion level during the initial learning phase. We made this change in order to minimize possible ceiling effects at the 20-min retention interval.

## Method

*Participants and design* A group of 184 individuals were randomly assigned to one of six groups (short or long test–feedback lag with a 20-min, one-day, or one-week retention interval). For the short-lag groups, 54, 32, and 22 people were randomly assigned to the 20-min, one-day, and one-week retention intervals. For the long-lag groups, 20, 23, and 33 people were randomly assigned to each retention interval group.[1]

*Materials* The items included 16 of the 24 face–name pairs (half male and half female) from Experiment 4. We reduced the number of items from Experiment 4 in order to decrease the chances of floor effects in the one-week retention interval, given that the criterion was changed from 70 % to 50 %.

*Procedure* The procedure was identical to that of Experiment 4, with the exception that the retention interval between the test–feedback phase and the final test phase occurred after 20 min, one day, or one week.

## Results and discussion

*Initial learning performance* The mean numbers of cycles to criterion were similar for the short- and long-lag schedules for the 20-min (3.76 vs. 2.95), one-day (2.66 vs. 3.26), and one-week (2.64 vs. 3.06) retention interval groups; none of the main effects were significant (all *p*s > .16), but the interaction was [$F(2, 178) = 3.87$, $p = .02$, $\eta^2 = .04$]. Due to the reliable interaction, the number of trials to criterion was used as a covariate in all subsequent analyses. Performance on the criterion trial did not differ for the short- and long-lag schedules for the 20-min (.63 vs. .67), one-day (.68 vs. .68), or one-week (.66 vs. .72) retention interval groups; none of the main effects nor the interaction were significant.

[1] The number of participants randomly assigned to each group was unexpectedly different in this experiment. Importantly, this was not due to differential attrition in each group since all participants had to return after a one week retention interval. Rather, it was because random assignment by the program assigned more participants to some groups than to others.

*Performance during the test–feedback and final test phases* Performance on the test trial during the test–feedback phase is presented in Table 1. As in all prior experiments, the results showed a significant main effect of learning schedule, with higher levels of test performance for the short-lag than for the long-lag group, $F(1, 177) = 64.26$, $p < .001$, $\eta^2 = .26$. The main effect of retention interval and the interaction were not significant (of course, retention interval was yet to be varied when these observations were made).

The critical data appear in Fig. 3, and show that we did replicate the predicted pattern of results by obtaining a lag effect at the one-week retention interval. However, we obtained no lag effect for either the 20-min or one-day retention interval groups. We had predicted that the lag effect would reverse at the 20-min retention interval, but it did not. A between-subjects ANOVA revealed a main effect of retention interval, with performance decreasing as the retention interval increased, $F(2, 177) = = 85.02$, $p < .001$, $\eta^2 = .46$. Although the interaction did not reach conventional levels of statistical significance ($F = 2.05$, $p = .131$), planned comparisons of the different retention intervals indicated no effect of lag at the 20-min ($F < 1$) and one-day ($F < 1$) retention intervals, but again a reliable effect at the one-week interval, $F(1, 52) = 5.01$, $p = .03$, $\eta^2 = .06$, replicating the results obtained in Experiment 4.

Finally, we evaluated recall for each of the short- and long-lag groups, as in the previous experiments. Table 1 shows performance on the test–feedback trial and final test trial for both the short- and long-lag groups. The results of a 2 (lag: short vs. long) × 2 (test type: test–feedback trial vs. final test trial) × 3 (retention interval: 20 min, one day, or one week) mixed-factor ANOVA showed a main effect of test type, with higher levels of performance on the test trial during the test–feedback phase than during the final test, $F(1, 177) = 7.88$, $p = .006$, $\eta^2 = .02$. The main effect of lag was significant, with higher levels of performance overall for the short-lag than for the long-lag group, $F(1, 177) = 12.56$, $p = .001$, $\eta^2 = .05$. Not surprisingly, the main effect of retention interval was also significant, with performance decreasing as the retention interval increased, $F(2, 177) = 29.01$, $p < .001$, $\eta^2 = .23$.

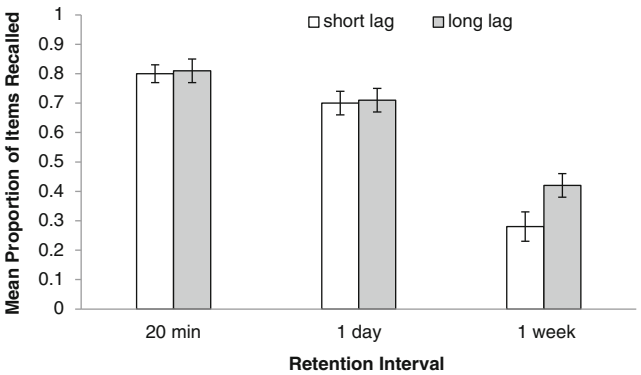


**Fig. 3** Mean proportions of items recalled on the final test trial as a function of lag and retention interval for Experiment 5. Error bars represent standard errors

Importantly, the interaction of test type and lag condition was once again significant, $F(1, 177) = 112.33$, $p < .001$, $\eta^2 = .10$, indicating different retention functions for short- and long-lag groups.

## General discussion

The primary focus of the present experiments was to examine the effects of test–feedback presentations when the lag between initial learning (to a criterion) and the test–feedback phase occurred across minutes or one day. Two important findings emerged from the seven experiments. First, when the final criterial test occurred one day after the test–feedback phase, no lag effect occurred (Exps. 1, 2a, 2b, 2c, 3, and select conditions in 5). Given the large literature that typically has shown lag effects with even short lags (e.g., 2, 4, 8, 20, and 40 intervening items; Madigan, 1969), failing to find a lag effect between our short (within minutes of learning items to criterion) and long (24 h after learning items to criterion) spacing schedules is notable. Second, when the retention interval was extended to one week, as opposed to one day, a robust lag effect emerged (Exps. 4 and 5). This outcome shows once again that lag effects depend on the amount of spacing and the retention interval used, as has been shown in work with much shorter lags (Balota et al., 1989; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). We see the same pattern in our much more extended paradigm, in that lag effects do not occur at "short" retention intervals (assuming, in this context, that a day is short), whereas the effect does emerge with a much longer retention interval (after a week).

These results are novel, and we will discuss each of them in turn.[2] First, we note that the effects of both testing and spacing have been shown to exert their influence more at long than at short retention intervals (e.g., Rawson & Kintsch, 2005; Roediger & Karpicke, 2006). For example, in the spacing-effect literature, Rawson and Kintsch had participants read expository texts twice. The spacing between reading the text was either short (occurred immediately after participants read the text the first time) or long (occurred one week later). A final recall test occurred immediately after the second study trial with the text (short retention interval) or two days later (long retention interval). Across two experiments, the results showed that the retention interval influenced the efficacy of the learning schedules. With a short retention interval, performance was greater for the short-spacing schedule. However, with the long retention interval, performance was greater for the long-spacing schedule. This research provides a conceptual replication of the experiments cited above with paired-associate materials and much shorter intervals (e.g., Balota et al., 1989). The same interaction occurs in the testing-effect literature: Repeated study (relative to a study and an initial test) produces better recall at a short retention interval, whereas the study–test sequence provides better performance at longer retention intervals (e.g., two days or one week in Roediger & Karpicke, 2006). The extension to these findings in our experiments is that a test–feedback trial included both test and restudy.

Interestingly, in the present experiments we found no difference in recall between the short- and long-lag schedules with relatively short retention intervals (20 min or one day), but we obtained a benefit for the long-lag schedule when the retention interval was increased to one week. It is surprising that we did not find a benefit for the short-lag schedule when the retention interval was 20 min relative to one day, given all of the prior work showing that performance is often greater for short (vs. long) spacing schedules when the retention interval is short (although we did change our procedure from a 70 % criterion in earlier experiments to a 50 % criterion in Exp. 5). This outcome indicates that terms like "short and long spacing" and "short and long retention interval" are relative; they do not depend on absolute time as the clock measures it, but on relative time in the context of the experimental conditions employed and the resulting event structure of the experiment. However, because of the criterion-learning procedure used in the initial acquisition phase before the spacing manipulation, participants always received multiple test–feedback events in each list that were spaced. Hence, this initial spacing may have overridden any influence of the spacing between the initial learning phase and the test–feedback phase, at least at retention intervals of 24 h or less. This particular pattern of results also provides additional evidence for the nonlinear relationship between lag and retention interval that has been documented elsewhere in the spacing-effect literature (see Delaney, Verkoeijen, & Spirgel, 2010, for a recent review).

In the present experiments, the test–feedback phase consisted of both a test and a restudy event, which precludes

[2] Because feedback was provided in the present experiments the results may have also reflected the contribution of test-potentiated learning (Arnold & McDermott, 2013a, 2013b; Izawa, 1966) when the spacing lag between original learning and the test–feedback phase was one day and the retention interval was one day. See Table 1 for the relevant comparisons that meet this criterion (24-h lag, 24-h retention interval). All experiments produced higher recall on the final test than on the test given 24 h previously (during the test–feedback practice phase), which was not observed in the short-lag conditions. Thus, the restudy after a test in the long-lag condition during the test–feedback practice phase provided a large boost to performance, which not only eliminated the expected forgetting over this one-day retention interval, but may have actually potentiated learning by showing an improvement. However, we refrain from favoring this explanation, because we did not have the proper control condition to demonstrate test-potentiated learning. Another difficulty in ascribing test-potentiated learning as a mechanism is that the initial learning phase in the present experiments was not a pure study phase, but instead an intermixed sequence of studying and testing (to criterion); this sequence occurred prior to the test–restudy sequence. Hence, the present design did not allow us to isolate the influence of test-potentiated learning, although this mechanism may have contributed to the observed patterns. Future research will be needed to pinpoint our improved recall after 24 h in the long-lag condition to test-potentiated learning.

our ability to have a clear understanding of the functional level of performance achieved at the end of the test–feedback phase. Thus, it is possible that the feedback during the test–feedback phase influences the two lag groups differently, which would make it difficult to determine whether lag or feedback (or the combination of the two) influenced the pattern of results presented in the present study (also see Footnote 2). We have another series of experiments (Pyc, Balota, McDermott, & Roediger, 2014) in which we evaluated the influence of lag (as in the present study), but we manipulated the type of practice that participants engaged in during the test–feedback practice phase. They either received repeated study (with no test), as in traditional spacing-effect experiments, or they received tests (with no feedback). Across a series of two experiments with 714 participants and varied materials (low-associate words pairs and face–name pairs), we found the same pattern of results reported in the present experiments. Specifically, with a one-day retention interval, we found no differences between short- and long-lag practice schedules. However, with a one-week retention interval, we found lag effects for both study and test groups. Thus, feedback is not solely responsible for the pattern of results found in the present study. Rather, the effect seems to be driven by the interaction of lag schedule and retention interval.

Our research also has implications for student learning, by showing that if a final criterial test occurs after a relatively short retention interval, then the particular schedule of test–restudy practice (at least within the present manipulation) does not much matter over the range of retention intervals up to 24 h. However, for students to retain information for longer periods of time (which we hope would be their goal), the long-lag schedule is clearly superior. This claim makes the assumption that our work with paired associates and criterion learning will generalize to more educationally relevant materials, which would need to be evaluated in future work. Additional research will be needed to evaluate how relearning episodes might influence the efficacy of spacing schedules (see Rawson & Dunlosky, 2011), and how various retention intervals might interact with the timing of relearning.

Despite the fact that we failed to find lag effects under conditions in which they might plausibly be expected to occur (i.e., with immediate vs. 24-h spacing of presentations and a 24-h retention interval), the present results can be interpreted within a number of theoretical accounts of spacing. For example, the desirable-difficulty framework (Bjork, 1994) posits greater benefits from a restudy episode when processing during restudy is more difficult relative to less difficult. "Difficulty of processing" during study can be defined in a number of ways, but in the context of the present experiments, the short-lag schedule would be akin to easier processing, and the long-lag schedule would be akin to more difficult processing. The results presented across seven experiments partially support the predictions of the desirable-difficulty framework. Specifically, recall was greater across a one-week retention interval when processing during the test–feedback trial was more difficult (the long-lag schedule) than when it was less difficult (the short-lag schedule). Recall increased between the test–feedback cycle and the final criterial test when the test–feedback cycle occurred 24 h after initial learning. This outcome occurred in all six experiments that included the relevant comparison. Importantly, however, the benefits of the longer lag on recall were not obtained until we used a one-week retention interval. The mystery is why the same effect did not occur after 24 h, a question to which we turn next.

The presence of the lag effect at the long but not the short retention interval may be best understood in terms of the distribution-based bifurcation model of retrieval practice effects advocated recently by R. A. Bjork and his colleagues (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). Although it is intended as an explanation for retrieval practice effects, the model may also be extended to accommodate our results, in which a test–feedback trial after a long lag (relative to a short lag) produces greater recall after a one-week retention interval, but does not with a one-day retention interval. For the present purposes, the most important implication of the model is its prediction that performance on a retention test will depend upon the interaction of the strength of learning during acquisition and the difficulty of the final test criterion. We unpack this logic in the next paragraph.

The bifurcation model assumes a normal distribution of items along a memory strength continuum before initial study. During study practice, all items are assumed to receive an incremental benefit in strength (shift along the distribution) with each new practice trial. In contrast, during test practice, items that are correctly retrieved will receive a greater shift along the distribution than during restudy practice, but items that are not retrieved will receive no benefit (because the failed retrieval attempt means that the item was not practiced). Hence, the distribution of item strengths becomes bifurcated. The difficulty of the final test criterion will determine which items in the distribution will be recallable on the final test. With a relatively easy final test criterion, many of the study items will be recalled because they have been sufficiently shifted along the strength dimension to permit recall. However, with a difficult final test criterion, only items that have moved farther along the distribution will be accessible for recall. These are the items that produce considerable benefit from successful retrieval practice during testing. The model accounts well for many effects found in the retrieval practice literature (e.g., no testing effect, or even a reverse testing effect, with a short retention interval, but a large testing effect with a long retention interval).

To explain our present set of results, we propose that different spacing schedules of test–feedback practice map on to strength distributions in a similar manner to studying and testing. Specifically, a short-lag schedule of test–feedback practice might show a pattern similar to that from simply studying items in the bifurcation model, because at a short interval almost all of them would be recalled. A long-lag schedule would show a pattern similar to that for tested items, because fewer items would be correctly retrieved, and those items would receive a greater incremental benefit in strength than would items recalled in the short-spacing schedule. Thus, in the long-lag case, the distribution would bifurcate because not all long-lag items would be retrieved. The difficulty of the final test would determine which items in the distribution would be recallable on that test. With a final test at relatively short retention intervals, many of the items in both the short- and long-lag test–restudy conditions would be recalled because they have been sufficiently shifted along the strength dimension to drive the recall test. However, with a difficult final test at a longer interval such as a week, only items that had moved farther along the distribution would be accessible.

We are assuming that retention interval is a proxy for "difficulty of the final test," which seems reasonable. However, as was pointed out above, what retention intervals are considered "short" versus "long" will depend on the conditions of the experiment. In the context of the present study, short retention intervals of 20 min or one day (Exps. 1–3 and 5) represent relatively easy final tests, relative to the long retention interval (a week; Exps. 4 and 5) providing a more difficult final test.

With a relatively easy final test (i.e., 20 min or one day), considerable overlap would exist in the items that could be recalled from the short- and long-lag schedules; thus, one may predict little influence of lag on final recall with the relatively short retention interval. However, with a more difficult final test (a week later), only items in the right tail of the shifted distribution would be accessible for recall, so one would expect to find a benefit for the long-lag schedule. The results reported in the present experiments are consistent with this account: With a short retention interval, we found no differences in performance on the final test for the short- and long-lag groups, but with a long retention interval, we obtained greater recall on the final test for the long-lag relative to the short-lag schedule.

One problem with our account is that we are rather arbitrarily using the terms "short" and "long." For lag effects, we refer to a few minutes as "short" and 24 h as "long"; for retention interval, we lump both 20-min and 24-h delays as "short" and distinguish the one-week interval as "long." As the astute reader will no doubt have noticed, we consider the same amount of time—24 h—to be long as a lag interval but short as a retention interval.

We must admit that the reasoning here is post hoc—we did not predict these effects, and indeed, the bifurcation model had not been published when we began these experiments—but the model does provide a way of understanding these results even if after the fact. (Post hoc is better than no hoc at all.) In addition, Bjork's desirable-difficulty and bifurcation model is also useful in understanding why test-potentiated learning is greater after a delay: The delay reduces the strength of the original encoding, and the test–feedback practice phase can then have a greater impact on the weakened representations, relative to when test–feedback practice occurs relatively soon after initial learning.

In summary, we found a lag effect between test–feedback and final test phases in our paradigm when the final retention interval was long (one week), but not when it was one day or shorter. The effects are robust, having been replicated with different materials as well as with variations to the paradigm. The finding of no lag effect (despite a 24-h lag) when the retention interval is 24 h is a surprise, given the relatively consistent findings of lag with even shorter intervals in other paradigms. The fact that a lag effect did occur under the same conditions after a one-week retention interval reveals (again) the dependence of lag effects on the retention interval used. We interpreted our results within the bifurcation model of Bjork and his colleagues, but further testing will be necessary. On a practical note, our results show that the spacing of practice at long intervals is definitely warranted for learning to be maintained over the long term (although the "long term" was only one week in our experiments).

**Author note** Supported by a grant from Dart Neuroscience, LLC. We thank David Blinn, Nicole McKay, John Slochower, Alexandra Taylor, and Teresa Yao for assistance with data collection and scoring.

## References

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review, 20,* 507–513. doi:10.3758/s13423-012-0370-3

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 940–945. doi:10.1037/a0029199

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4,* 316–321. doi:10.1111/j.1467-9280.1993.tb00571.x

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag and retention interval. *Psychology and Aging, 4,* 3–9.

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

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

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,* 354–380. doi:10.1037/0033-2909.132.3.354

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

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale: Erlbaum.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 53, pp. 63–147). San Diego: Academic Press.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84,* 795–805.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.)*. New York: Columbia University, Teachers College. Original work published 1885.

Glenberg, A. M., & Lehmann, T. S. (1980). Spacing repetitions over 1 week. *Memory & Cognition, 8,* 528–538. doi:10.3758/BF03213772

Goverover, Y., Basso, M., Wood, H., Chiaravalloti, N., & DeLuca, J. (2011). Examining the benefits of combining two learning strategies on recall of functional information in persons with multiple sclerosis. *Multiple Sclerosis Journal, 17,* 1488–1497.

Greene, R. L. (2008). Repetition and spacing effects. In H. L. Roediger (Ed.), *Learning and memory: A comprehensive reference* (Cognitive psychology of memory, Vol. 2, pp. 65–78). Oxford: Elsevier.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19,* 126–134. doi:10.3758/s13423-011-0181-y

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports, 18,* 879–919.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1250–1257. doi:10.1037/a0023436

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. doi:10.3758/BF03194055

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25,* 498–503.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97.

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory, 20*, 37–47.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Litman, L. & Davachi, L. (2008). Distributed learning enhances relational memory consolidation. *Learning and Memory, 15*, 711–716.

Maddox, G. B., & Balota, D. A. (2012). Self control of when and how much to test face–name pairs in a novel spaced retrieval paradigm: An examination of age-related differences. *Aging, Neuropsychology, and Cognition, 19,* 620–643. doi:10.1080/13825585.2011.640658

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 828–835.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23. doi:10.3758/s13428-011-0124-6

Melton, A. W. (1967). Repetition and retrieval from memory. *Science, 158,* 532. doi:10.1126/science.158.3800.532-b

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 596–606.

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology, 66,* 206–209. doi:10.1037/h0046694

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., Balota, D. A., McDermott, K. B., & Roediger, H. L., III. (2014). *Is There a Benefit of a 24 Hour Spacing Interval? No After a Day, Yes After a Week*. Manuscript in preparation.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97,* 70–80.

Robbins, D., & Bush, C. T. (1973). Memory in great apes. *Journal of Experimental Psychology, 97,* 344–348.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Sargis, E. G., Skitka, L. J., & McKeever, W. (2013). The Internet as psychological laboratory revisited: Best practices, challenges, and solutions. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding our online behavior* (2nd ed., pp. 253–270). Oxford: Oxford University Press.

Simone, P. M., Bell, M. C., & Cepeda, N. J. (2012). Diminished but not forgotten: Effects of aging on magnitude of spacing effect benefits. *Journals of Gerontology, 68B,* 674–680. doi:10.1093/geronb/gbs096

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology, 25,* 763–767.

Toppino, T. C., Fearnow-Kenney, M. D., Kiepert, M. H., & Teremula, A. C. (2009). The spacing effect in intentional and incidental free recall by children and adults: Limits on the automaticity hypothesis. *Memory & Cognition, 37,* 316–325. doi:10.3758/MC.37.3.316

Tully, T., Preat, T., Boynton, S. C., & Del Vecchio, M. (1994). Genetic dissection of consolidated memory in Drosophila. *Cell, 79,* 35–47.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20,* 568–579.

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16,* 80–84. doi:10.1111/j.1467-8721.2007.00480.x

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133,* 273–293. doi:10.1037/0033-2909.133.2.273