Running head: RETRIEVAL PRACTICE OVER THE LONG TERM

This manuscript was accepted for publication in
Psychonomic Bulletin & Review on March 11, 2014.

Retrieval Practice Over the Long Term:

Should Spacing Be Expanding or Equal-Interval?

Sean H.K. Kang[1], Robert V. Lindsey[2], Michael C. Mozer[2], & Harold Pashler[1]

[1]University of California, San Diego

[2]University of Colorado, Boulder

Author Note

Sean H.K. Kang, Department of Psychology, University of California, San Diego; Robert V. Lindsey, Department of Computer Science, University of Colorado, Boulder; Michael C. Mozer, Department of Computer Science, University of Colorado, Boulder; Harold Pashler, Department of Psychology, University of California, San Diego. Sean Kang is now at Department of Education, Dartmouth College.

This work was supported by the Institute of Education Sciences (Grant R305B070537 to H. Pashler), NSF (Grant BCS-0720375, H. Pashler, PI; and Grant SBE-0542013, G.W. Cottrell, PI), and a collaborative activity award from the J.S. McDonnell Foundation. The second author was supported by an NSF Graduate Fellowship. We thank Noriko Coburn for assistance.

Correspondence should be addressed to Sean Kang, Department of Education, Dartmouth College, Hanover, New Hampshire 03755. Email: sean.kang@dartmouth.edu

Abstract

If there are multiple opportunities to review to-be-learned material, should a review occur soon after initial study and recur at progressively expanding intervals, or should the reviews occur at equal intervals? Landauer and Bjork (1978) argued for the superiority of expanding intervals, whereas more recent research has often failed to find any advantage. However, these prior studies have generally compared expanding versus equal-interval training within a *single session*, and assessed effects only *upon a single final test*. We argue that a more generally important goal is to maintain high average performance over a considerable period of training. For the learning of foreign vocabulary spread over four weeks, we found that expanding retrieval practice (sessions separated by an increasing number of days) produced equivalent recall to equal-interval practice on a final test given eight weeks after training. However, the expanding schedule yielded much higher average recallability over the whole training period.

It is well established in the memory literature that reviews spaced apart in time enhance long-term retention of material more than reviews that occur soon after initial study (*the spacing effect*; see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a survey) and that reviews are more effective when they involve testing instead of re-presentation (also known as *retrieval practice*; e.g., Carrier & Pashler, 1992; Kang, McDermott, & Roediger, 2007; see Roediger & Karpicke, 2006, for a review). Spacing and testing can be combined—i.e., spaced retrieval practice—to obtain the benefits of both.

Two broad theories of the spacing effect have been influential. According to *encoding variability theory,* the degree of match in the context at encoding and the context at retrieval determines the probability of successful retrieval (i.e., context serves as a retrieval cue); increasing the time interval between initial study and review heightens the difference in contextual elements that are encoded at both instances, and thereby increases the likelihood of the encoded contexts overlapping with the context at the final test administered after a delay (e.g., Glenberg, 1976). *Study-phase retrieval theory* provides an alternative account: Final memory performance benefits from restudy to the extent that the second encounter with an item reminds the learner of the previous encounter (i.e., an automatic study-phase retrieval; Thios & D'Agostino, 1976; Wahlheim, Maddox, & Jacoby, 2014); also, the benefit is greater the more effortful the study-phase retrieval is, which explains the advantage of spacing (Pyc & Rawson, 2009).

The bulk of research on spaced retrieval practice has focused on how the lag between an initial study episode and a single review opportunity affects performance on a later final test (e.g., Landauer & Eldridge, 1967; Cepeda et al., 2009). In many real-word contexts, however, learners have more than one opportunity to review the to-be-remembered material, in which case the relevant question is how these multiple reviews should be distributed over time in order to optimize learning.

Expanding Retrieval Practice

Landauer and Bjork (1978) were the first to compare the efficacy of various schedules of retrieval practice. In one experiment, subjects studied first name-last name pairs once, followed by a practice phase in which they made three attempts to retrieve the appropriate last name when cued with a first name (no feedback was provided). In the *massed* condition, the three retrieval attempts occurred consecutively right after the initial presentation of the pair; in the *equal-interval* (spaced) condition, the number of intervening items between each retrieval attempt was kept constant; in the *expanding* condition, the first retrieval attempt occurred soon after the initial presentation, followed by a progressively larger number of intervening items between successive retrieval attempts; in the *contracting* condition, retrieval was attempted only after a relatively large number of intervening items, followed by fewer and fewer intervening items between successive retrieval attempts. On a final test given shortly after practice, an expanding schedule of practice yielded the highest recall (followed by equal-interval, contracting, and massed practice, respectively). Based on these results, Landauer and Bjork argued for the superiority of expanding retrieval as a form of spaced practice. Their explanation was that attempting retrieval soon after initial presentation of the item insures a high level of success, and since successful retrieval strengthens the learning of the item, subsequent retrieval attempts can be progressively delayed without compromising the level of success, while still maintaining the effectiveness of each subsequent retrieval in strengthening memory for the item (Bjork & Bjork, 1992). The idea of expanding retrieval practice is intuitively appealing (see Leitner, 1972, for a similar proposal with flashcards), and has become influential as a technique for training both normal individuals (e.g., Metzler-Baddeley & Baddeley, 2009) and individuals with cognitive impairments (e.g., Camp, Bird, & Cherry, 2000).

However, many recent experimental reports have questioned whether expanding interval training is really superior to equal-interval practice (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Carpenter & DeLosh, 2005; Karpicke & Roediger, 2010). Indeed, several recent studies have even found the opposite result when using a delayed final test (i.e., retention interval of at least a day; Cull, 2000; Logan & Balota, 2008). For example, Karpicke and Roediger (2007) found that while an expanding schedule of practice yielded better performance than equal-interval practice on an immediate final test (replicating Landauer & Bjork’s, 1978, original findings), the pattern reversed on a delayed final test (given 2 days after training). They suggested that the placement of the first retrieval attempt was more important than the relative spacing of subsequent retrieval attempts in determining long-term retention. To maximize long-term retention, that first retrieval attempt needs to be challenging or effortful (i.e., occurring after some delay rather than immediately after initial presentation of the item), and in the view of these authors, this may be why equal-interval practice trumps expanding practice at longer retention intervals.

More recently, Storm, Bjork, and Storm (2010) demonstrated that whether expanding or equal-interval retrieval practice was superior in a particular situation depended critically on the rate of forgetting of the to-be-remembered material: When forgetting was rapid (due to the presentation of intervening information that was highly interfering), expanding practice produced better recall on a delayed final test than equal-interval practice (see also Maddox, Balota, Coane, & Duchek, 2011).

Limitations of Previous Research

Although studies comparing expanding and equal-interval retrieval practice have revealed intriguing interactions between type of schedule and other variables (e.g., forgetting rate of the material, whether feedback is provided during training), the practical relevance of these findings is limited due to two factors. First, in this research, type of schedule has virtually always been manipulated within a single learning session (an exception is Cull, 2000, Experiments 3 & 4), with spacing operationalized in terms of the number of intervening items between successive repetitions of a target item. But practice within any single session, however it may be scheduled, is rarely adequate to support long-term retention.

Second, previous research has focused solely on optimizing performance on a single final test. By contrast, in most real-world learning scenarios (e.g., acquiring a foreign language or on-the-job training), the learned material should be accessible over a long period of time, and the paradigm of a training period followed by a single test may be irrelevant. Instead, the training and test periods may be confounded in a single period of time, and material should be reviewed within this window so as to ensure or maximize the continuous accessibility of the material. That is, instead of optimizing study for a single test in the future, reviews should be scheduled to maximize the average recall performance in the training period.

The two limitations of previous work that we mention—the short time scale of experiments and the focus on a final test—are related, because when the time scale of training is short and items are practiced multiple times within a single session, the recallability of material between retrieval attempts is irrelevant, but in naturalistic learning scenarios which operate over a much longer time scale, the recallability of material between study sessions may be more important than the recallability following the end of the study period.

Present Study

Our experiment was conducted over a time scale adequate to have relevance to education and training: The training period was 28 days, with a final test administered 56 days later. Subjects were presented with 60 Japanese-English word pairs to learn. After initial study followed by 3 cycles of retrieval practice for all items on Day 1, items assigned to the expanding condition underwent additional retrieval practice on Days 3, 9, and 28, whereas items assigned to the equal-interval condition underwent additional practice on Days 10, 19, and 28. Corrective feedback was provided during retrieval practice (as in most real-world training, but unlike many laboratory studies).

To evaluate the continuous accessibility of material during the training period, one would ideally like to inject tests throughout the training period. However, because of the contamination that these tests can cause, it would be necessary to remove items once they have been tested, and such a procedure would therefore require a very large participant and/or item population, and the protocol would impose strong demands on participants. As an alternative, we probed memory only infrequently during the training period, and used memory models to estimate levels of recall and forgetting between probes.

Method

Subjects

Subjects were recruited from our laboratory's internet subject pool. 37 subjects with no prior knowledge of Japanese completed all 7 sessions of the experiment for $35 payment. The mean age of the subjects who completed the experiment was 36.4 years (range: 20–63 years), and 22% were male.

Stimuli

The study material consisted of 60 Japanese-English word pairs.[1] For each subject, 20 items each were randomly assigned to the expanding and equal-interval retrieval practice conditions, and the remaining 20 served as filler items (for use in fitting parameters of a model that we will describe later). The filler items were studied on Day 1, and half were tested a single time on Day 9 and the other half were tested a single time on Day 28.

Design and Procedure

Schedule of training was manipulated within-subjects. In the expanding condition, items received additional retrieval practice on Days 3, 9, and 28. In the equal-interval condition, items received additional retrieval practice on Days 10, 19, and 28.

In the first session (Day 1), subjects first were presented with all the Japanese-English word pairs once (in a random order), for 8 s each, with a 1-s blank screen after each item. After initial presentation of the items, there was a 30-s distractor task (counting backward by 3s), followed by 3 cycles of retrieval practice for all items. The order of items on each practice cycle was randomized, with the constraint that the first 2 items of each cycle would not be the last 2 items in the previous cycle. On each retrieval practice trial, the Japanese word would first be presented alone for 6 s, and during that time subjects were asked to retrieve and type in the English equivalent if they could. After 6 s had elapsed, the intact Japanese-English pair would be presented for 2 s (regardless of how the subject responded), followed by a 1-s blank screen.

Subjects were reminded via emails and were given a 24-h window (starting at 12 h before the appointed time and ending at 12 h after the appointed time) to log in for subsequent sessions. Subjects that missed the time window for any of the sessions were dropped from the experiment. Sessions 2 to 6 consisted of 3 cycles of retrieval practice for the items assigned to practice on that day/session. For Session 6 (Day 28), the items from the expanding and equal interval conditions were randomly intermixed during retrieval practice.

For the final session (Day 84), subjects received a final test on the items. The test trials were self-paced—the Japanese words were presented singly and subjects could take as much time as they needed to type in the English equivalent. No feedback was provided. After completing the experiment, subjects were debriefed and thanked for their participation.

Results

Mean recall proportions during the training phase and on the final test as a function of training schedule are displayed in Figure 1. The figure shows performance during each of the three cycles of retrieval practice that occurred in each training session.

Note that all items were studied once prior to retrieval practice on Day 1, and that corrective feedback was provided after each retrieval practice trial in each session.

Training Phase

Performance at the beginning of training was very similar across the expanding and equal-interval conditions: The level of recall during the third cycle of retrieval practice on Day 1 was not different between the two conditions (.30 vs. .29), $t(36) < 1$, suggesting that the items randomly assigned to both conditions were of equivalent difficulty. At the end of training, performance also seemed fairly similar across conditions. During the third cycle of retrieval practice on Day 28, the proportion of items recalled was not reliably different between the expanding and equal interval conditions (.62 vs. .65), $t(36) = 1.429$, $p = .162$.

Final Test Performance

The expanding condition yielded numerically higher recall than the equal interval condition on the final test (.49 vs. .46), but this difference was not statistically reliable, $t(36) = 1.23$, $p = .227$. In terms of amount of forgetting between the end of training and the final test (i.e., the difference in recall between the third cycle of retrieval practice on Day 28 and recall on the final test), the expanding condition resulted in significantly less forgetting than the equal interval condition (.13 vs. .19), $t(36) = 2.321$, $p = .026$, $d = 0.38$.

Assessing Recallability Over the Training Period

Despite the seeming parity in performance across training conditions at the start and at the end of training, the question we began with was: if participants were probed at a random time during the training period, what would their average recall level over the entire period be? To measure this directly would require an impractical experiment in which participants would be probed at very fine intervals throughout the training period. Although we did not do that, the data collected allows us to estimate accessibility of the learned information while relying only on well-grounded and fairly minimal assumptions about the learning and forgetting processes.

We assume that forgetting between sessions follows a generalized power function (Wixted & Carpenter, 2007). Because items are practiced three times within a session, we know that the final practice trial (a test followed by feedback/study) should boost recall higher than the level of performance observed on the test itself, but we do not know precisely how much. A conservative heuristic used for estimating the gain from the final practice trials within each session is described in the Supplementary Materials. Given this estimate of recall proportion at the end of a session, along with recall proportion at the first test of the next session, two constraints are imposed on the forgetting function. Because the generalized power-law forgetting function has three parameters, two constraints are insufficient, yielding some residual uncertainty about the shape of the forgetting function. In Figure 2, we represent this uncertainty by sampling 250 curves that are consistent with the initial and final points of the forgetting function. The faint, thin lines represent these samples. The solid line superimposed over each set of faint lines is the expectation of the samples. The sampling and fitting procedures are described in detail in the Supplementary Materials.

The conclusions are quite clear, as seen in Figure 2: The area under the expanding-interval curve is greater than the area under the equal-interval curve. Quantitative measures are consistent with the visual impression: The mean estimated recall proportion over the Day 1–28 period is .51 for the expanding condition but only .43

for the equal-interval condition. This difference is reliable when treating the sampled forgetting functions as the random variable, $t(498) = 83.7$, $p < .0001$.[2]

Arguably a better measure of reliability is to treat subjects as the random variable. We interpolated forgetting curves for each subject using the methodology described in the Supplementary Materials, and once again found a reliable improvement for the expanding over the equal-interval condition (.49 vs .41, $t(36)=3.65$, $p<.001$). Further statistical analysis confirming that the expanding condition yielded greater average accessibility can be found in the Supplementary Materials.

Discussion

Spaced retrieval practice has been shown to benefit long-term retention, but there has been a long-running debate over the best way to schedule or distribute the retrieval attempts when there are multiple opportunities to practice retrieval. Two contenders have emerged: In an expanding schedule, retrieval is attempted soon after initial study, followed by subsequent retrieval attempts that occur after progressively longer delays; in an equal-interval schedule the first retrieval attempt occurs only after some delay, and the interval between successive retrieval attempts is uniform. Proponents of expanding schedules have argued that these insure successful retrieval on the first attempt, which strengthens the memory, and in turn allows for successive retrieval attempts to occur at longer and longer delays, thus maximizing the memory enhancement of each retrieval opportunity (Landauer & Bjork, 1978). But several studies have found an expanding schedule to be inferior to equally spaced practice when retention is assessed after a long delay, and some critics have suggested that having the first retrieval occur so soon after initial study obviates the benefits of retrieval, in essence causing that retrieval attempt to be wasted (Karpicke & Roediger, 2007).

While these previous studies have uncovered factors that may modulate the relative effectiveness of expanding versus equal-interval schedules, it was argued above that they are rather limited in practical relevance, due to generally having training confined within only a single session. Spaced retrieval practice has obvious applications in the fields of education and training (e.g., Dempster, 1991), but it is unclear whether the existing findings from the laboratory generalize at all to cases in which review of the material occurs over a longer period of time. In addition, prior research has focused primarily on criterial performance on a final test. But in the context of training that is spread out over a long span of time, it is as important—if not more so—to consider performance *during* training as a metric of efficacy.

The present experiment examined the relatively efficacy of expanding and equal-interval retrieval practice for the learning and retention of foreign vocabulary, with retrieval practice occurring in sessions that were separated by days (over a span of 4 weeks). When considering the average amount of information that was accessible over the training phase, practice with an expanding schedule was clearly advantageous. Moreover, when memory was assessed 8 weeks after the last session of training, recall performance was not worse (and actually slightly better) in the expanding than equal interval condition. The final test data assures us that the more rapid acquisition in the expanding condition was not accompanied by more rapid forgetting (cf. Karpicke & Roediger, 2007; Logan & Balota, 2008).

Our findings suggest that when retrieval practice is spread out over days or weeks, scheduling the review sessions in an expanding fashion produces better average performance then equal-interval spacing over the training period. Expanding practice not only produces faster acquisition and greater access to the material over the training period, it was even observed to slightly retard forgetting over the long term too.

Prevailing theories of spaced practice have generally not focused directly on the maintenance of information in memory (i.e., resistance of the memory trace to interference or forgetting; Küpper-Tetzel & Erdfelder, 2012). For instance, encoding variability theory focuses on retrieval processes—overlaps in encoding and retrieval contexts serve as effective retrieval cues. Study-phase retrieval theory, on the other hand, focuses on encoding processes—an optimal lag between initial study and review is one that yields effortful but successful study-phase retrieval, which leads to superior re-encoding of the information. Thus, these theories do not make specific predictions about the schedule of spaced practice that would produce superior accessibility to the information that is being learned over a lengthy training period. The present study was not designed to adjudicate between theories of spaced practice, but the results seem especially congenial to the idea of study-phase retrieval benefiting learning. The advantage of the expanding schedule can, in part, be explained by the early review sessions occurring relatively soon after initial study (yet separated by days, thus allowing a higher probability of successful but effortful retrievals than the early review sessions in the equal-interval practice condition), accompanied by the later review sessions spread relatively farther apart (to foster effortful retrieval). The results of the present study, however, are not entirely consistent with encoding variability theory: The close spacing of early sessions would not seem to be optimal for the encoding of material in diverse contexts. The *multiscale context model* (MCM; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009), a computational model of memory accessibility that incorporates assumptions of encoding variability, study-phase retrieval, as well as predictive utility (Staddon, Chelaru, & Higa, 2002), provides a good fit for the present data (see Supplementary Materials for more details).

Future research might profitably examine a number of questions. While the differences in overall performance found here are sizable, the overall level of performance is rather low. It will be interesting to see if the advantage of the expanding schedule remains when people are trained to a high criterion of success in the initial session. Another important question is whether the present findings scale up to time periods of years instead of months. Given that spacing effects with two sessions have been found to scale up with increases in the time intervals involved (Cepeda et al., 2009), it seems plausible that they would—but establishing this point will require additional empirical work.

References

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

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Camp, C.J., Bird, M.J., & Cherry, K.E. (2000). Retrieval strategies as a rehabilitation aid for cognitive loss in pathological aging. In R.D. Hill, L. Backman, A. Neely Stigsdotter (Eds.), *Cognitive rehabilitation in old age* (pp. 224–248). Oxford, UK: Oxford University Press.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

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,* 236–246.

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.

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

Dempster, F.N. (1991). Synthesis of research on reviews and tests. *Educational Leadership, 48,* 71–76.

Glenberg, A.M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15,* 1–16.

Kang, S.H.K., McDermott, K.B., Roediger, H.L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558.

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

Karpicke, J.D., & Roediger, H.L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38,* 116–124.

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.

Landauer, T.K., & Eldridge, L. (1967). Effect of tests without feedback and presentation-test interval in paired-associate learning. *Journal of Experimental Psychology, 75,* 290–298.

Leitner, S. (1972). *So lernt man lernen.* Freiburg im Breisgau, Germany: Herder.

Logan, J.M., & Balota, D.A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15,* 257–280.

Maddox, G.B., Balota, D.A., Coane, J.H., & Duchek, J.M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology & Aging, 26,* 661–670.

Metzler-Baddeley, C., & Baddeley, R.J. (2009). Does adaptive training work? *Applied Cognitive Psychology, 23,* 254–266.

Mozer, M.C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1321–1329). La Jolla, CA: NIPS Foundation.

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.

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

Storm, B.C., Bjork, R.A., & Storm, J.C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38,* 244–253.

Thios, S.J., & D'Agostino, P.R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*, 529–536.

Wahlheim, C.N., Maddox, G.B., & Jacoby, L.L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 94–105.

Wixted, J.T., & Carpenter, S.K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, *18*, 133–134.

Footnote

[1] The Japanese words were in their Romanized form, and were selected from a set of Japanese-English word pairs provided to us by Philip Pavlik.

[2] In response to a reviewer's query about the average recallability over the *entire* experiment (not just during the training period), we performed the same analysis using recall data from Days 1–84. The estimated mean accessibility was 52.3% and 48.7% in the expanding and equal-interval conditions, respectively. The advantage of expanding over equal-interval practice, though smaller, was still statistically significant, $t(498) = 45.3$, $p < .0001$. While this result is interesting and informative, we should point out that such an analysis, strictly speaking, no longer compares average recallability *during* expanding and equal-interval practice (e.g., practice sessions were not distributed equally across the entire 84-day period).

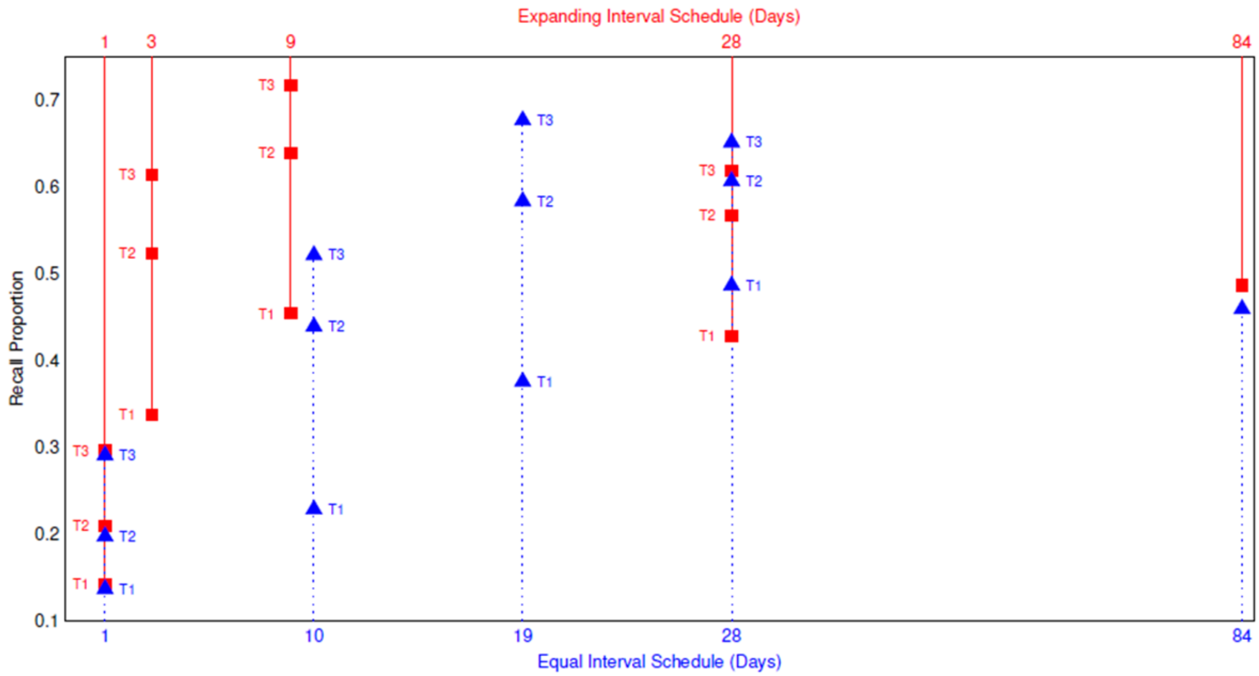

Figure 1. Mean recall proportion over the course of the experiment for the expanding-and equal-interval schedule. T1, T2, and T3 refer to the first, second, and third test (retrieval practice) cycles, respectively, during each session of the training phase. The days on which items in the expanding condition were practiced are shown along the top of the graph, and are connected to the recall proportions for that condition (squares) by a solid line dropping down from the top of the graph. The days on which items in the equal-interval conditions were practiced are shown along the bottom of the graph, and are connected to the recall proportions for that condition (triangles) by a dashed line rising up from the bottom of the graph.

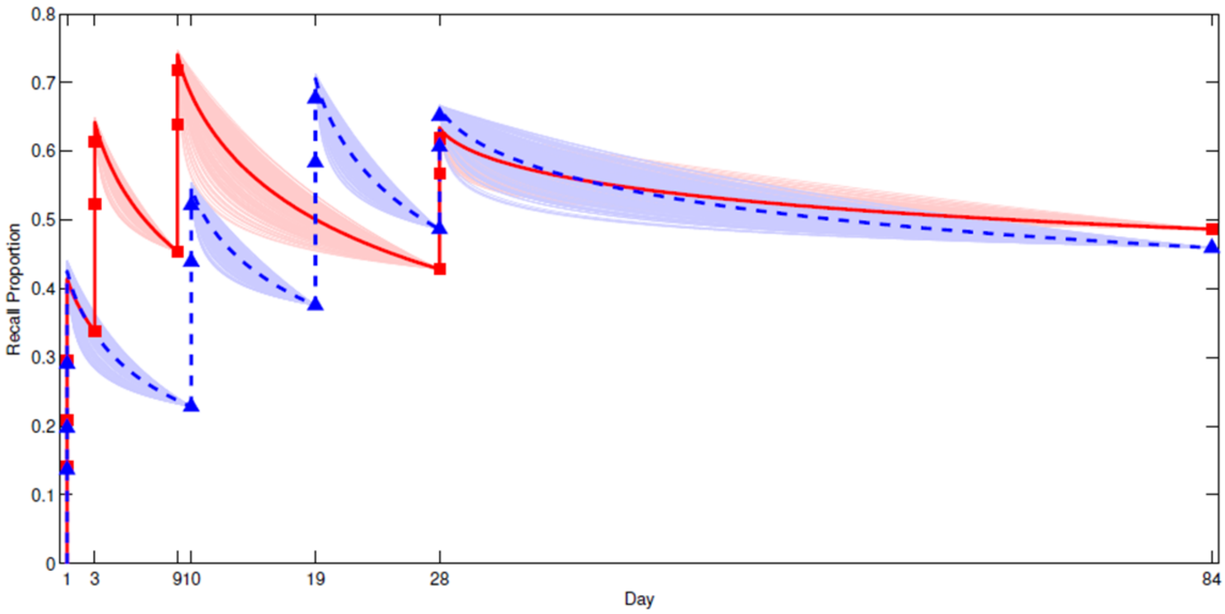

Figure 2. Interpolation of recall performance over the entire experiment. The training phase lasted from Days 1 to 28. Expanding and equal-interval conditions are depicted by solid and dashed lines, respectively. The squares and triangles indicate observed mean recall performance in the expanding and equal-interval conditions, respectively. As explained in the text, the faint lines represent uncertainty in the shape of the forgetting curves.

Supplementary Materials

The procedure for estimating the recallability curves presented in Figure 2 involved two steps: (1) estimating memory accessibility at the end of a practice session, and (2) estimating the shape of the forgetting curve between practice sessions.

Each session involved three trials in which an item was tested and then studied further (feedback). Consequently, although we probed the state of memory during a session, we do not know its state at the end of a session—after the final round of study. However, we can extrapolate from the three tests to estimate the end-of-session recallability. For example, if the three tests (T1, T2, T3) yield proportions correct of .1, .2, and .3, we might imagine that following the final study, the recall proportion would be .4.

We observed an interesting regularity involving the three tests. If p_i denotes the proportion correct on test *i*, then we can define a measure of improvement due to the study following test *i* as the proportion increase in recall:

$$Improvement_i = \frac{p_{i+1} - p_i}{p_i}. \tag{1}$$

The improvement from T2 to T3, relative to the improvement from T1 to T2,

$$ImprovementRatio = \frac{Improvement_2}{Improvement_1}, \tag{2}$$

was roughly constant across the various sessions after the initial day of study, suggesting to us that we might predict the improvement from T3 to (a hypothetical) T4 by assuming the ratio is a fixed constant, leading to:

$$Improvement_3 = ImprovementRatio\ x\ Improvement_2 \tag{3}$$

and combining Equations (1)-(3), we obtain

$$p_4 = (1 + Improvement_2^2 /\ Improvement_1)p_3.$$

This prediction of performance at the end of a session, which looked quite reasonable on visual inspection (see starting point of forgetting curves in Figure 2, positioned above T3), provides an initial point on a forgetting curve, and a final point is simply the performance on the first test of the next session. We fit these two points to a generalized power function,

$$r = \alpha(1 + \beta t)^{-\gamma},$$

which characterizes retrieval probability *r* as a function of the time elapsed since study, *t*. Because this function is underconstrained by the two data points, a family of solutions for parameters $\{\alpha, \beta, \gamma\}$ exists. We sampled 250 instances from this family using a nonlinear least squares curve fitting function in MATLAB with different random initialization points, and obtained the faint lines in Figure 2. The mean of this set is depicted by the dark lines in Figure 2. Because of the uncertainty in the value of p_4, we also jittered its value in the resampling process.

This same procedure was used both to estimate the mean memory state across participants (Figure 2) and the memory state of individual participants. The latter estimates were used for evaluating the statistical reliability of the difference between the expanding and equal-interval conditions.

For further evidence that the expanding condition yields greater average accessibility of the material, we fit individual subject data to a model of memory specifically designed to predict the strength of memory following multiple spaced practice sessions. This model, the *multiscale context model* (MCM; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009), is fit to data points from a forgetting curve characterizing memory strength as a function of time following a single practice session. MCM then predicts memory strength continuously over time following one or more practice sessions. The data used to constrain MCM were 5 points along the forgetting curve collected in the course of the experiment (recall tests with retention intervals of 0, 2, 8, 9, and 27 days). MCM predicts memory strength curves that are quite close to those in Figure 2, and matches the data in Figure 2 despite the fact that most of the data points in the figure were not used for constraining model parameters. More relevant for the present purpose, the model predicts that the mean recall proportion in the training period (Day 1 to 28) is larger for the expanding than equal-interval condition (.46 vs. .40, $t(36) = 10.38$, $p < .001$).

Reference

Mozer, M.C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1321–1329). La Jolla, CA: NIPS Foundation.