Running head: RETRIEVAL PRACTICE OVER THE LONG-TERM

# Retrieval Practice Over the Long Term:

# Expanding or Equal-Interval Spacing?

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.

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 is supported by an NSF Graduate Fellowship. We thank Noriko Coburn for assistance.

Correspondence should be addressed to Sean Kang, Department of Psychology, University of California, San Diego, La Jolla, California 92093-0109. Email: seankang@ucsd.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 4 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 8 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; for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1996) 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). 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. (The expanding, contracting, and equal-interval conditions were matched in terms of average spacing between 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, but yet 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; but see Cull, Shaughnessy, & Zechmeister, 1996). 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 in a series of experiments on the learning of Graduate Record Exam vocabulary 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. In summarizing the extant literature, Balota, Duchek, and Logan (2007) concluded that "the additional benefits of expanded practice over equal interval practice have not been well substantiated in recent research." (p. 100) 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, in press).

### *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). In other words, spacing in almost all studies to date have been operationalized in terms of the number of intervening items between successive repetitions of a target item within a continuous sequence of training trials. 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. In 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*

We present a study that addresses the two limitations of previous work. Our experiment was conducted over a time scale sufficiently long for educational and occupational relevance: The training period was 28 days. To evaluate the effect of training on long-term retention, a final test was administered 56 days later. Subjects were presented with 60 Japanese foreign vocabulary 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 (similar to most training in the real world).

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 research subject pool, which includes people of various ages and from various countries, screened for careful attention to directions, English proficiency, and conscientious participation in prior studies. None reported having any prior knowledge of Japanese. Subjects who completed the experiment received $35 payment in the form of Amazon.com gift certificates.

We report data from subjects who completed all 7 sessions of the present experiment within the specified time windows (*N* = 37). 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. In other words, subjects were tested and given feedback, in a procedure modeled after Carrier and Pashler (1992).

Subjects were reminded via emails to log in for subsequent sessions. They were given a 24-h window (starting at 12 h before the appointed time and ending at 12 h after the appointed time). For example, subjects were allowed to start the second session (Day 3) between 36 and 60 h after the start of the first session. 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 appropriate items (i.e., items assigned to the condition that involved practice on that day/session). The procedure for each retrieval practice trial was identical to that described earlier. Again, the order of items was randomized for each cycle, with the constraint that the first 2 items in a cycle would not be the last 2 items in the previous cycle. 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, we would like to determine the accessibility of the material over the extended training period. In other words, if participants were probed at a random time during the training period, choosing all times with equal probability, what would average recall be? To answer this question definitively, we require data in which participants are probed at fine intervals over the training period. The difficulty of obtaining these data is obvious. However, due to the testing procedure within each session of retrieval practice, we do have some information about the state of memory within the training period. Further, we can make reasonable assumptions concerning the nature of forgetting between sessions. Taken together, we can interpolate the state of memory over the entire training period.

Figure 2 shows such an interpolation. Forgetting between sessions is assumed to follow a generalized power function (Wixted & Carpenter, 2007). Because items are practiced three times within a session, we know that the final practice trial (i.e., test followed by feedback/study) should boost recall higher, but we do not have an empirical assessment of how much higher. The Appendix describes a heuristic for estimating the benefit in recall probability of the final practice trial within a session. Given an estimate of recall proportion at the end of a session, along with recall proportion at the first test of the next session, we have two constraints on the forgetting function. Because the generalized power-law forgetting function has three parameters, two constraints are insufficient, and therefore there is some uncertainty in 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 Appendix describes the sampling and fitting procedure in detail.

On inspection, Figure 2 suggests that the area under the expanding-interval curve is greater than the area under the equal-interval curve: Between Days 3 and 19, material is more recallable in the expanding condition, and between Days 19 and 28, material is more recallable in the equal-interval condition. Quantitative measures are consistent with the visual impression: Mean recall proportion over the Day 1–28 period is .51 in the expanding condition but only .43 in the equal-interval condition. This difference is reliable when treating the sampled forgetting functions as the random variable, $t(498) = 83.7$, $p < .0001$.

Perhaps a more useful measure of reliability is to treat subjects as the random variable. We interpolated forgetting curves for each subject using the methodology described in the Appendix, and once again found a reliable improvement for the expanding condition over the equal-interval condition (.49 vs .41, $t(36)=3.65$, $p<.001$). For further evidence that the expanding condition yields greater average availability 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 is larger for the expanding than equal-interval condition (.46 vs. .40, $t(36) = 10.38$, $p < .001$).

## Discussion

Spaced retrieval practice has been shown to benefit long-term retention, but there is a 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, on the other hand, 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 rather than over equal intervals produces better overall performance over the training period. Expanding practice not only produces faster acquisition and greater access to the material over the training period, it seems to retard forgetting over the long term too.

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 each training session (as one would do whenever the marginal cost of extra training time is small relative to the cost of arranging for a training session). Doing this will probably cause choice of schedule to become confounded with differences in total study time, and thus would not allow one to draw conclusions about efficiency (as is possible in the current study), but it might better approximate the choices that present themselves in real-world settings. 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 an empirical test of this question would be worthwhile.

# References

Balota, D.A., Duchek, J.M., & Logan, J.M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J.S. Nairne (Ed.), *The foundations of remembering* (pp. 83–105). New York: Psychology Press.

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., Shaughnessy, JJ., & Zechmeister, E.B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied, 2,* 365–378.

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.

Dempster, F.N. (1996). Distributing and managing the conditions of encoding and practice. In E.L. Bjork & R.A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 317–344). San Diego, CA: Academic Press.

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.

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. (in press). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology & Aging.*

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.

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.

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

# Appendix

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 expanding and equal-interval conditions.

# 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.

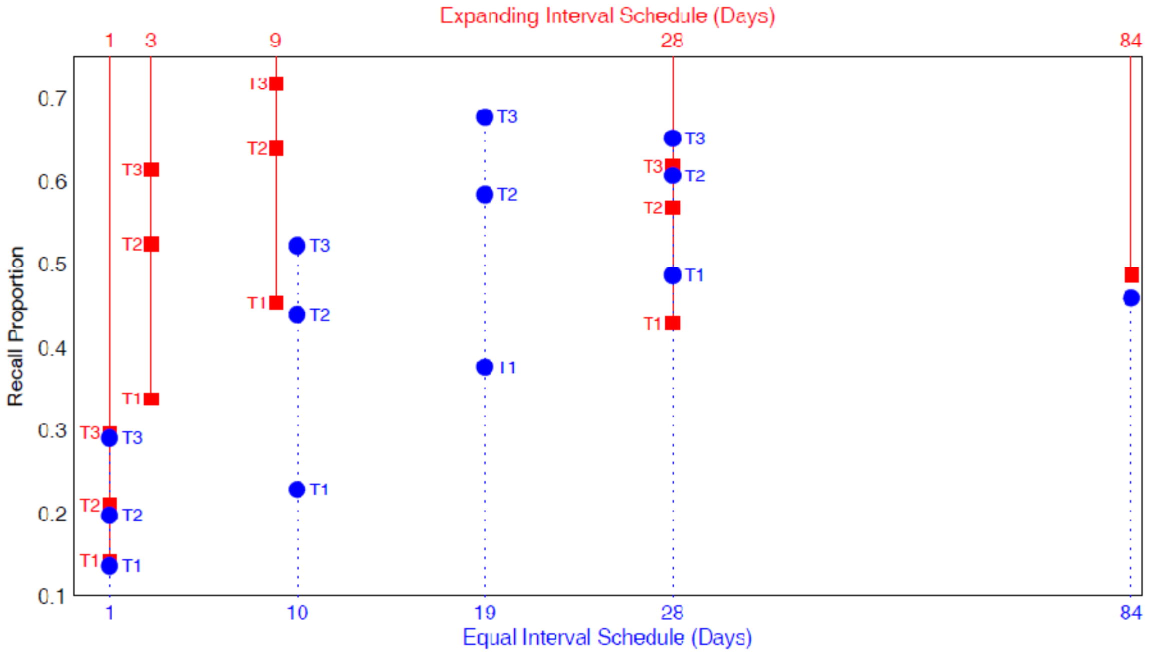

*Figure 1*. Mean recall proportion over the course of the experiment for the expanding-and equal-interval schedule. T1, T2, and T3 refer to the first, second, and third test (retrieval practice) cycles, respectively, during each session of the training phase. The days on which items in the expanding 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 (circles) by a dashed line rising up from the bottom of the graph.

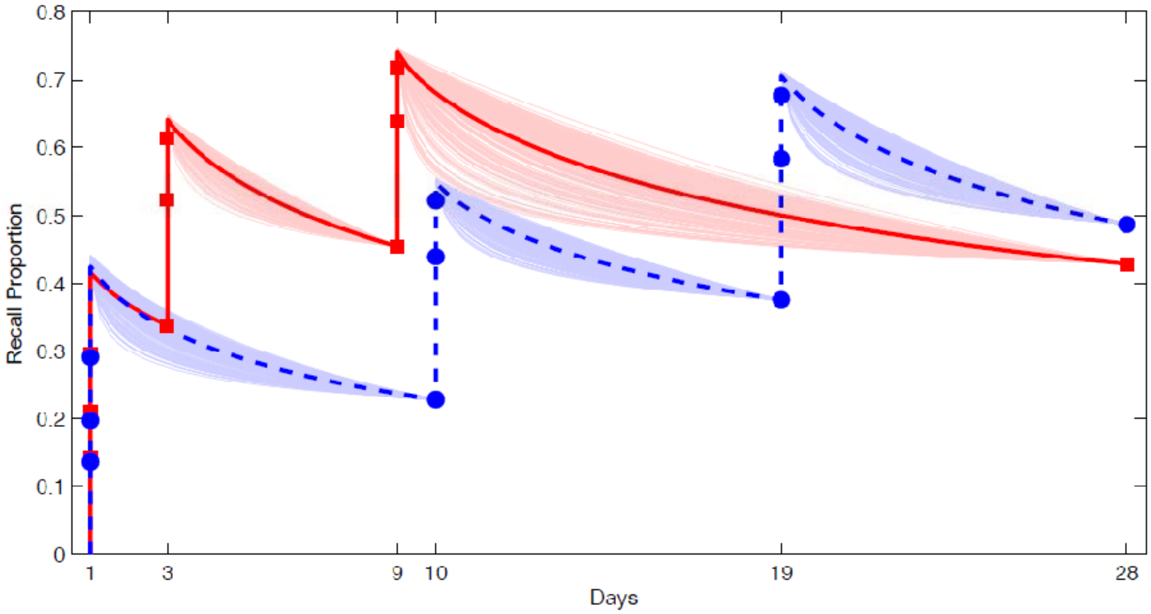

*Figure 2.* Interpolation of recall performance in the training phase of the experiment. Expanding and equal-interval conditions are depicted by solid and dashed lines, respectively. The squares and circles indicate observed mean recall performance in the expanding and equal-interval conditions, respectively. As explained in the text, the faint lines represent uncertainty in the shape of the forgetting curves.