(in press version): This manuscript may differ from the final published version

Running Head: OPTIMIZING RETENTION USING DISTRIBUTED PRACTICE

Spacing effects in learning: A temporal ridgeline of optimal retention

Nicholas J. Cepeda
York University and University of California, San Diego

Edward Vul
University of California, San Diego and Massachusetts Institute of Technology

Doug Rohrer
University of South Florida

John T. Wixted and Harold Pashler
University of California, San Diego

To achieve enduring retention, people must usually study information on multiple occasions. How does the timing of study events affect retention? Prior research has examined this issue only in a spotty fashion, usually with very short time intervals. To characterize spacing effects over significant durations, over 1350 individuals were taught a set of facts and – after a gap of up to 3.5 months – given a review on the same facts. A final test was administered at a further delay of up to 1 year. At any given retention interval, an increase in the inter-study gap at first increased, and then gradually reduced, test performance. The optimum gap value was about 20% of the test delay for delays of a few weeks, falling to about 5% when delay was one year. The interaction of gap and test delay implies that many educational practices are likely to be highly inefficient.

Keywords: spacing effect, distributed practice, long-term memory, instructional design

Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention

As time progresses, people lose their ability to recall past experiences. The amount of information lost per unit of time gradually shrinks, producing the well-known increasingly gradual forgetting curve. Far less is known about the course of forgetting after a person has experienced *multiple* exposures to the same piece of information. Multiple exposures are obviously very common, and are probably essential for most long-term instruction. Thus, an understanding of how the gap between two exposures affects subsequent forgetting is fundamental to any effort to temporally structure learning events in a rational manner. This could have important benefits if it turns out that gap has big effects on recall—as the data described here demonstrate—and an analysis of the issue should also help in constraining theories of the processes underlying long-term memory.

Effects of gap on later memory are usually termed “distributed practice” or “spacing” effects, and there is a large literature on such effects going back to the 19th century (for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Crowder, 1976; Dempster, 1988). A spacing experiment should involve multiple periods of study devoted to the same material, separated by some variable time *gap*, with a final memory test administered after an additional *retention interval* (RI) measured from the second exposure (Figure 1). Many spacing studies have shown that no gap results in worse final test performance than does a brief gap. Several studies involving modest time intervals ranging from minutes to days have found that memory at the final test is best for intermediate gap durations (e.g., Balota, Duchek, & Paullin, 1989; Glenberg, 1976; Glenberg & Lehmann, 1980; Young, 1966; see Cepeda et al., 2006, for a meta-analysis focused on this point).

Given the enormous size of the literature on spacing effects, the reader may wonder why there would be a need for further and more systematic exploration. Indeed, the literature is large: A recent review of distributed practice studies involving verbal recall (Cepeda et al., 2006) examined more than 400 papers. However, only about a dozen of these looked at retention intervals as long as one day, with just a handful examining retention intervals longer than one week. While psychologists have decried the lack of practical application of the spacing effect (Dempster, 1988; Rohrer & Taylor, 2006), the fault appears to lie at least partly in the research literature itself: Based on short term studies, one cannot answer with confidence even such basic questions as "How much time between study sessions is appropriate to promote learning and retention over substantial time intervals—is it a matter of days, weeks, or months?"

In one pioneering study involving long retention intervals (Bahrick et al., 1993), acquisition and retention of foreign language vocabulary were examined over several years using four subjects. In this study, the subjects were trained to a fixed performance criterion within each study session (as they were in Bahrick & Phelps, 1987). The results showed that increasing the inter-study spacing out to 56 days improved performance (Figure 2).

The Bahrick studies might appear to suggest that over these long intervals, spacing effects may be monotonic in character, rather than showing an inverted-U shape, as noted in shorter-term studies. However, 56 days is actually a relatively short proportion of the extremely long retention intervals used in Bahrick et al.'s study, and it is this ratio that is probably most critical, as will emerge below.

Another issue in interpreting the Bahrick studies (Bahrick et al., 1993; Bahrick & Phelps, 1987) is the use of a fixed performance criterion. Given the forgetting that takes place during the gap between study sessions, this inevitably means that many more relearning trials will be necessary during most sessions at the longest gap (as compared to the shortest). While one might argue that in some cases students will wish to relearn to criterion, and thus the Bahrick procedure may be informative about the appropriate timing of such a relearning session, nonetheless this design feature makes it challenging to draw conclusions about the efficiency of study because it confounds total study time and spacing gap in favor of greater spacing gaps.

The goal of the present study was to examine the joint effects of gap and retention interval more systematically and over more substantial time intervals than has been done previously. We held constant the number of restudy trials in the second study session, thus allowing us to look at the effect of gap apart from the amount of time provided for restudy. Furthermore, by including a much greater range of gap/RI ratios, we aimed to assess the generality of the possibly non-monotonic relationship of retention to gap, and more generally, to reveal something about the shape of what we will term the *retention surface,* that is, final test performance as a function of gap and RI.

This undertaking requires running thousands of total subject training and test sessions. Fortunately, the advent of internet-based experimental testing panels makes it feasible to carry out multiple learning and test sessions with a very large number of individuals on a remote basis. As described in the Appendix, the reliability and validity of internet data collection has become increasingly clear in recent years. While objections are still occasionally raised against this form of data collection, these receive little support from actual experience with the method.

Preliminary Data

In a preliminary laboratory-based study (Cepeda et al., 2007) that provides a key benchmark for the present study, 150 subjects participated in three sessions over a period of up to one year. The first two sessions were learning sessions in which the subjects were taught a set of obscure but true facts (e.g., *Snow golf was invented by Rudyard Kipling*) and the names of some obscure visually presented objects (e.g., *coccolith*). These two study sessions were separated by a gap ranging from 10 minutes to 6 months. All subjects then returned to the lab for a final memory test 6 months after their second learning session. The non-monotonic pattern noted in short-term studies was indeed found: Recall success (for both facts and names) was best for a 1-month gap, with much worse recall for shorter gaps and slightly poorer recall for longer ones. If the optimal gap value should happen to increase linearly with the retention interval, then these results would imply that about a 20% ratio of gap to retention interval optimizes retention, but no such thing can be observed.

Current Study

We now report a more comprehensive set of learning episodes and tests involving 1354 new subjects from our laboratory's Internet Memory Research panel formed for long-term repeat testing, in what we suspect may be the most systematic analysis of long-term spacing effects yet carried out. To properly characterize the interaction of gap and retention interval, the current study combines various gap and RI values, for a total of 26 different conditions. In the first learning session, 32 facts were learned to a criterion of one perfect recall. After the prescribed gap, a second learning session was performed. Here, subjects were tested twice on each fact, and after responding, were shown the correct answer. After a retention interval, subjects were given two tests on each of the 32 facts, without feedback. The first was a recall test (*Who invented snow golf?*), and the second was a recognition test in which subjects tried to pick the correct answer from among 5 equally-likely alternatives.

Method

Subjects.

Subjects were drawn from our laboratory's online research subject pool, which includes subjects of various ages living in a wide variety of countries. Each time they participate in study, subjects are entered into a drawing for cash prizes.

We report data from subjects who completed all three sessions within the necessary time windows. Non-completion rates increased at the longer delays, as one would expect in any multi-year study (Table 1), but subjects who completed all three sessions did not show any reliable differences in their initial test performance from those who did not complete the final test. There were also no detectable differences between the groups in age, gender, number of obscure facts known before beginning the study, or a wide array of background and demographic characteristics. Mean age of subjects was 34 years (*SD* = 11; range = 18 – 72), and 72% were female.

Stimuli and materials.

Stimuli consisted of 32 obscure but true trivia facts (e.g., "What European nation consumes the most spicy Mexican food?" Answer: "Norway"). All answers consisted of a single word of 5 or 6 letters. As shown in Table 1, there were a range of gaps (interval between sessions 1 and 2) and retention intervals (interval between the second learning session and the final test). *Design and procedure.*

There were 26 gap-by-retention interval combinations, and each subject was randomly assigned to one of these. The number of gaps for each retention interval (RI) varied: 6 gaps for each of 2 RIs, and 7 gaps for each of the other 2 RIs. Retention intervals and gaps were chosen so that there were five gaps in common to all the retention intervals, and to ensure that each retention interval was associated with gaps that produced gap/RI ratios near 0.1, 0.2, and 0.3.

This experiment was conducted on a web server running the open source LAMP (Linux, Apache, MySQL, PHP) framework. The study was programmed in HTML, PHP, and JavaScript, and subjects could access the experiment from any standard web browser.

We assigned a disproportionately large number of subjects to conditions requiring longer total time between sessions, in order to compensate for the anticipated greater non-completion rates for those groups. In the first session, subjects were told that they would be tested on a series of facts, with feedback. Each fact was presented in question form; the subject was encouraged to guess if they were not confident of the answer, and then correct-answer feedback was provided. The first presentation of each fact allowed us to identify and remove from analysis any items known to the subject prior to the study. Questions answered correctly on the very first test were assumed to be known by the subject and were excluded from all subsequent analyses for that subject only. Subjects were trained to a criterion of successfully answering each of the 32 questions correctly, cycling through the list of items not yet answered correctly. Whenever a question was answered correctly, it did not appear again in the first training session. Each cycle involved a new random ordering of the list of items. Subjects answered between 61 and 96 questions in the course of the first session before they reached criterion. Subjects were advised by email when it was time for them to perform the second session. When gap was nominally zero, the second session began without any delay after the first (the actual length of the zero-day gap was about 3 min, or .00256 days).

In the second learning session, the same entire list of questions was run through twice, in a different random order, with each item being followed by a presentation of the correct answer. The subject could take as long as he or she wished to answer, or leave the item blank. Regardless of the subject’s response, the correct answer was then displayed for 4 seconds, and the next question appeared after approximately 1 second.

During the final session, subjects were given two tests, each covering all 32 facts. No feedback was provided in this phase. The first was a recall test. The second was an easier multiple choice recognition test, which offered five potential answers to each question (e.g., “What European nation consumes the most spicy Mexican food? (a) Norway; (b) France; (c) Poland; (d) Spain; (e) Greece"). Each of the five alternative answers was chosen about equally often by a separate pilot sample who had not been exposed to the facts.

Results

Figure 3 shows the effect of gap on recall and recognition for each of the four different retention intervals, for the subjects who completed all the phases of the study. For each retention interval, final performance always rose initially with increasing gap and then fell as gap was further increased. The effects of gap were very large in magnitude: For a fixed amount of study time, the optimal gap provided a 64% increase, *d* = 1.1, in final recall, and a 26% increase, *d* = 1.5, in final recognition, as compared to a zero-day gap (in the present article, d values refer to the comparison of zero-day versus optimal gap). For the RIs of 7, 35, 70, and 350 days, the optimal gaps (of those included in the study) were 1, 11, 21, and 21 days, respectively, for recall data, and 1, 7, 7, and 21 days, respectively, for recognition data. All of these findings are in good agreement with the lab benchmark dataset, which, with its RI of 6 months, lay about where it would be expected to between the RI = 70 and RI = 350 conditions—despite the fact that the laboratory study fixed session 2 exposure duration and the current study did not. (Both studies provided a fixed number of relearning trials.)

For 7, 35, 70, and 350 day RIs, percentage improvements in recall for optimal versus zero-day gaps were 10, 59, 111, and 77%—$t(124) = 6.5$, $p_{\text{rep}} = .99$, $d = 1.3$; $t(111) = 8.9$, $p_{\text{rep}} = .99$, $d = 0.6$; $t(102) = 8.6$, $p_{\text{rep}} = .99$, $d = 1.7$; $t(68) = 3.9$, $p_{\text{rep}} = .99$, $d = 0.9$, respectively. For recognition data, percentage improvements for optimal versus zero-day gaps were 1, 10, 31, and 60%—$t(124) = 2.3$, $p_{\text{rep}} = .99$, $d = 0.7$; $t(136) = 7.5$, $p_{\text{rep}} = .99$, $d =$

1.5; $t(112) = 8.7$, $p_{\text{rep}} = .99$, $d = 1.7$; $t(68) = 7.9$, $p_{\text{rep}} = .99$, $d = 2.1$ for 7, 35, 70, and 350 day RIs, respectively. Estimated optimal gap values rise as retention interval is increased, and these depart noticeably from the fixed ratio of retention interval suggested by some earlier researchers based on much shorter-term studies (Crowder, 1976; Murray, 1983).

Discussion

The results presented here document the existence of enormous and non-monotonic spacing effects that unfold over very long periods of time, with study time equated across conditions. As noted earlier, performance on the final test can be represented as a retention surface in which performance is plotted as a function of study gap and retention interval. One such function that provides a good fit to our data ($R^2 = .98$) is shown in Figure 4, and this function satisfies four constraints conveyed by our data. First, for any gap duration, recall performance must decline as a function of RI (i.e., test delay) in a negatively accelerated fashion in order to produce the familiar forgetting curve consistent with more than 100 years of memory findings. Second, for any RI greater than zero, an increase in study gap should cause recall to first increase and then decrease. Third, as RI increases, the optimal gap should increase (see Figure 3A), as shown by the direction of the red ridgeline in Figure 4. Fourth, as RI increases, the ratio of optimal gap to RI must decline. In Figure 4, for example, the optimal gap for RI = 350 equals 23 days, which is just 7% of the RI.

The surface in Figure 4 is an instance of the general form,

$$\text{Recall} = A\,(bt + 1)^{-R},$$

where A equals immediate recall performance (i.e., when t = 0), *R* equals the rate of forgetting, and *b* is a temporal scaling parameter (cf. Wixted, 2004). Initial recall performance (A) varies with gap according to the function,

$$A = p + (1 - p)\,e^{-a\,g},$$

where *p* and *a* are parameters. This function ensures that an increase in gap causes immediate recall performance to decline from perfection (when g = 0) to an asymptote equal to *p*. The rate of forgetting (R) also varies with gap according to the function,

$$R = 1 + c\,(\ln{(g + 1)} - d)^2,$$

where *c* and *d* are parameters. This is a U-shaped function of the natural log of study gap, meaning that, for each test delay (*t*), increasing the study gap causes the rate of forgetting to drop quickly before increasing more slowly thereafter. (The surface in Figure 4 has the following parameter values: p = 0.760, a = 0.017, b = 0.011, c = 0.092, and d = 3.453.) Exploration of the data indicated that a number of other functions can also provide quite decent fits to the data, with various tradeoffs between interpretability of parameters and simplicity of the function, and we do not contend that this function offers a uniquely accurate characterization of the surface—merely a reasonable one.

Theoretical Implications

The overall shape of this surface seen over such long intervals may help in constraining mechanisms of the spacing effect. Theories that attribute effects of gap to a reduced likelihood of information residing in short-term memory, such as most forms of deficient processing theory (Jacoby, 1978; Rundus, 1971), would not seem to fit well with present data (although this mechanism might operate under other conditions, of course). Working memory operates on a time scale of seconds or minutes, whereas gap effects are seen on a scale of day and weeks (optimal gap is several weeks, for longer retention intervals). All-or-none theories (Estes, Hopkins, & Crothers, 1960), in which items are either learned or not learned on any given trial, may also be challenged by the present data. This theory suggests that spacing will benefit learning when the first learning episode has been forgotten; thus, longer study gaps should always produce better final-test retention, and there should not be an optimal gap. Other distributed practice theories, such as encoding variability (Glenberg, 1979) and study-phase retrieval (Murray, 1983) are potentially consistent with the basic results seen in Figure 3.

Recent simulation work in our lab suggests that some recent quantitative theories (Pavlik & Anderson, 2003; Raaijmakers, 2003) may have trouble accounting for the present data, especially when these accounts are forced to explain not only final-test data, but also the performance seen in Session 2 (Mozer, Cepeda, Pashler, Wixted, & Rohrer, 2007). We have conducted our own simulations of Pavlik and Anderson's ACT-R model and Raaijmaker's SAM model, in order to determine if these models can characterize the ridgeline of optimal retention. We were not able to fit both the increase in optimal gap as a function of RI and the decrease in gap/RI ratio as a function of increasing RI. Whether or not this conclusion stands, it seems likely that the data provide significant new constraints on theorizing about spacing over meaningful time intervals.

Educational Implications

The present results show that the timing of learning sessions can have powerful effects on retention with study time equated—effects that, as with our benchmark study, seem far larger than what are typically seen in short-term spacing studies (Cepeda et al., 2006). However, for practical purposes, the results also reveal a sobering fact: The optimally efficient gap between study sessions is not some absolute quantity that can be recommended, but a quantity that depends dramatically upon the retention interval (a point that was evident in the short-term studies such as Glenberg, 1976, and is now seen to extend to far greater time intervals). To put it simply, if you want to know the optimal distribution of study time, you need to decide how long you wish to remember something. Although this poses challenges for practical application, certain conclusions can nonetheless be drawn. If a person wishes to retain information for several years, a delayed review of at least several *months* seems likely to produce a highly favorable return on a time investment—potentially doubling the amount ultimately remembered, holding study time constant—as compared to less temporally distributed study. While in agreement with the earlier work of Bahrick, this advice it at odds with many conventional educational practices—for example, study of a single topic confined within a given week of a course. Based on the current results, this compression of learning into a too-short period is likely to produce misleadingly high levels of immediate mastery that will not survive the passage of substantial periods of time (supporting the arguments of Bahrick, 2005, Dempster, 1988, and Schmidt & Bjork, 1992). It is also of interest to note that while there are costs to using a gap that is longer than the optimal value, these costs are much smaller than the costs of using too small a gap value, as evidenced by the fact that, as gap increases, accuracy increases steeply and then declines much more gradually (Figure 3). In light of the present results, it appears no longer premature for psychologists to offer some rough practical guidelines to those who wish to use study time in the most efficient way possible to promote long-term retention.

References

Bahrick, H. P. (2005). The long-term neglect of long-term memory: Reasons and remedies. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Decade of behavior* (pp. 89-100). Washington, DC: American Psychological Association.

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

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.

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.

Birnbaum, M. (1999). Testing critical properties of decision making on the internet. *Psychological Science*, *10*, 399–407.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2007). *Optimizing distributed practice: Theoretical analysis and practical implications*. Submitted for publication.

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.

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

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627-634.

Estes, W. K., Hopkins, B. L., & Crothers, E. J. (1960). All-or-none and conservation effects in the learning and retention of paired associates. *Journal of Experimental Psychology*, *60*, 329-339.

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.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*, 95-112.

Glenberg, A. M., & Lehmann, T. S. (1980). Spacing repetitions over 1 week. *Memory and Cognition, 8*, 528-538.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust
web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*, 93-104.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649-667.

Krantz, J. H., and Dalal, R. (2000). Validity of web-based psychological research. In M. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 35-60). San Diego, CA: Academic Press.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science, 11,* 502–506.

Mozer, M. C., Cepeda, N. J., Pashler, H., Wixted, J. T., & Rohrer, D. (2007). Temporal and associative context variability: An encoding variability model of distributed practice. Manuscript in preparation.

Murray, J. T. (1983). Spacing phenomena in human memory: A study-phase retrieval interpretation (Doctoral dissertation, University of California, Los Angeles, 1982). *Dissertation Abstracts International*, *43*, 3058.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187-193.

Pavlik, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Doerner, & H. Schaub (Eds.), *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 177-182). Bamberg, Germany: Universitats-Verlag Bamberg.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, *27*, 431-452.

Reips, U-D. (2002). Standards for internet experimenting. *Experimental Psychology, 49*, 243-256.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209-1224.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63-77.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207-217.

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review, 111*, 864-879.

Young, J. L. (1966). Effects of intervals between reinforcements and test trials in paired-associate learning (Doctoral dissertation, Stanford University, 1966). *Dissertation Abstracts International*, *27*, 3699.

Author Note

Nicholas J. Cepeda, Department of Psychology, York University, and Department of Psychology, University of California, San Diego; Edward Vul, Department of Psychology, University of California, San Diego, and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; Doug Rohrer, Department of Psychology, University of South Florida; John T. Wixted and Harold Pashler, Department of Psychology, University of California, San Diego.

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H020061 and R305H040108 to H. Pashler, and by the National Science Foundation (Grant BCS-0720375, H. Pashler, PI; Grant SBE-0542013, G. W. Cottrell, PI). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Harry Bahrick and an anonymous reviewer for helpful comments on this manuscript.

Correspondence concerning this article should be addressed to Hal Pashler, Dept of Psychology, University of California, San Diego, USA; hpashler@ucsd.edu

Appendix: Comments and Validity of Internet Memory Testing Data

Internet testing has become common in the behavioral sciences over the past several years, and standards have now been developed based on early experiences with this method. Our internet testing procedures follow the standards recommended by Reips (2002). The validity of internet testing has been well supported in recent reviews (Gosling et al., 2004), with excellent correspondence between results obtained with internet samples are results obtained in the laboratory (e.g., Birnbaum, 1999; Krantz & Dalal, 2000; McGraw, Tew, & Williams, 2000; Reips, 2002). This tracks our own experience in conducting memory studies in the lab and on the web. This correspondence is again seen in comparing the present internet-based results with the laboratory benchmark data discussed in the text. Speaking more informally, it is the authors' impression that the average level of care and caution shown by our internet panel actually tends to exceed that shown by the typical undergraduate fulfilling an experiment participation requirement mandated for a class.

However, several objections are sometimes raised against internet testing, and these deserve comment. One potential objection is that internet subjects may have more distractions than subjects tested in a laboratory. However, a comparison of the distribution of overall memory performance scores found with internet samples (including the present one) does not suggest any meaningful difference. For example, comparing gap and RI values that were roughly the same in both the current study and the benchmark study, average performance on the final test was 41% in the current study and 45% for the benchmark study.

Another potential concern sometimes raised is the possibility of "cheating" (i.e., writing down answers). Note that due to the randomized between-subject design used here, even if there were some small incidence of cheating and/or more severe distraction than occurs in the lab, these elements would have merely introduced noise, thus dampening the effects of the temporal variables, which–as we have seen–were large in magnitude and corresponded well to those obtained in laboratory studies. Moreover, in examining the distribution of overall memory performance, we saw little evidence of suspiciously good performance. The proportion of subjects whose performance might be termed "surprisingly good" (arbitrarily defined as 85% correct or better on the final test) was 2.6% for the internet sample as compared to 2.1% for comparable gap by RI values in the lab benchmark study. The lack of evidence for cheating is not surprising, given that subjects were explicitly asked not to engage in such behavior, along with the fact that there were no incentives favoring it.

In summary, while it is understandable for researchers to view web-collected data with initial caution, actual experience with these methods provides little reason to view web-collected data as any less credible than lab-collected data.

Table 1: Number of Subjects in Each Experimental Condition

RI (days)	Gap (days)	Number of Subjects
7	0	60
7	1	66
7	2	79
7	7	77
7	21	70
7	105	45
35	0	72
35	1	69
35	4	75
35	7	66
35	11	41
35	21	61
35	105	23
70	0	55
70	1	67
70	7	59
70	14	51
70	21	49
70	105	27
350	0	45
350	1	34
350	7	43
350	21	25
350	35	41
350	70	26
350	105	28

Figure Captions

Figure 1. Structure of a typical study of spacing effects on learning. Study episodes are separated by a varying gap, and the final study episode and test are separated by a fixed retention interval.

Figure 2. Final test performance as a function of gap, in data reported by Bahrick et al. (1993) involving spacing over multi-year retention intervals. For each of the four values of retention interval used in their experiment, accuracy increased monotonically as gap increased. However, the largest gap by RI value was only 15% (56 days / 365 days), which (in light of data from the present study) probably accounts for the failure to demonstrate the non-monotonicity seen in Figures 3 and 4.

Figure 3. Performance on the recall (a) and recognition (b) tests as a function of gap, for each of the four retention intervals. Points are mean accuracy +/- 1 *SEM*. Lines correspond to cubic spline fits to the data, with fixed points at gaps of 0 and 105 days. Optimal gap increases with increasing retention interval, and there is a non-monotonic gap effect at each RI.

Figure 4. A functional approximation of recall on the final test (as a proportion), plotted as a function of gap and test delay (i.e., retention interval). The red ridgeline is comprised of points representing the optimal performance for each test delay. The forgetting function for each gap is a power function. The location of the ridgeline indicates that, as test delay increases, optimal gap increases while there is a decrease in the ratio of optimal gap to test delay. See text for parameter values and a fuller description of this surface.

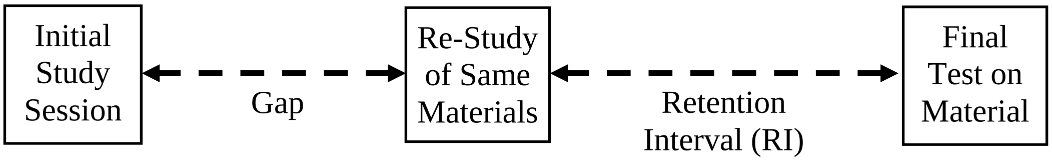

Figure 1

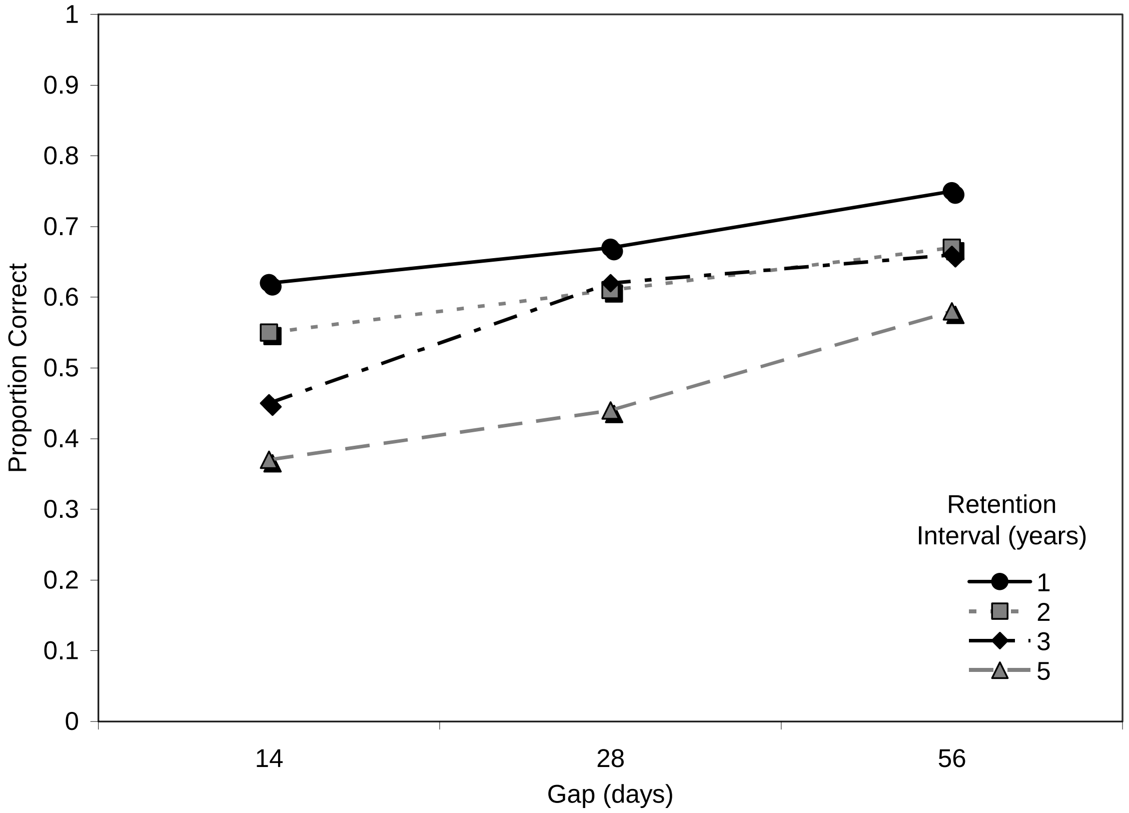
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion Correct
14
28
56
Gap (days)
Retention
Interval (years)
1
2
3
5

Figure 2

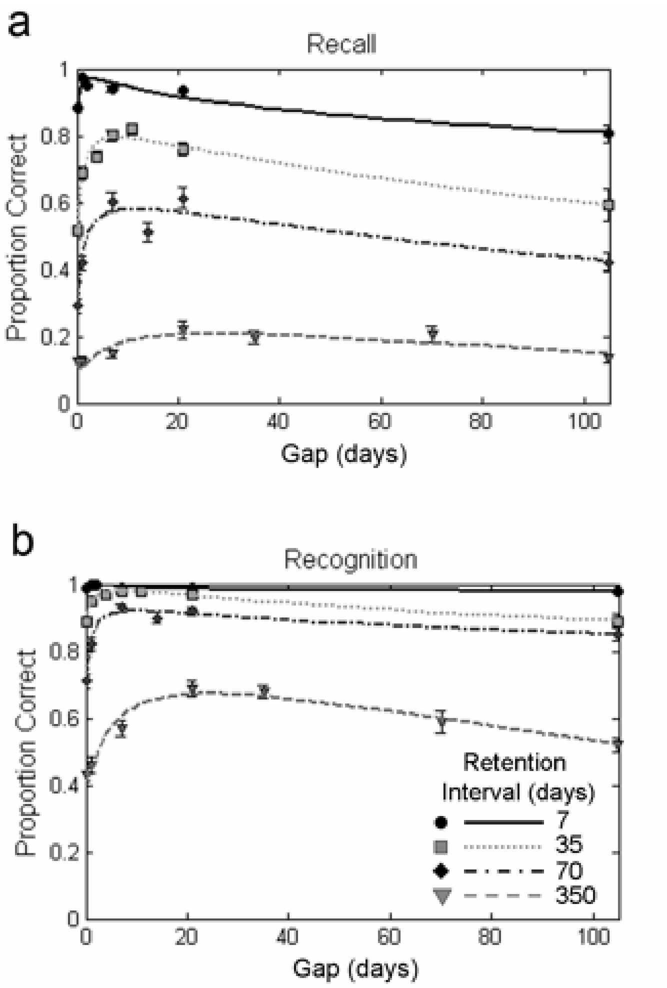
a
Recall
Proportion Correct
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
Gap (days)
b
Recognition
Proportion Correct
Retention
Interval (days)
● 7
■ 35
◆ 70
▼ 350
Gap (days)

Figure 3

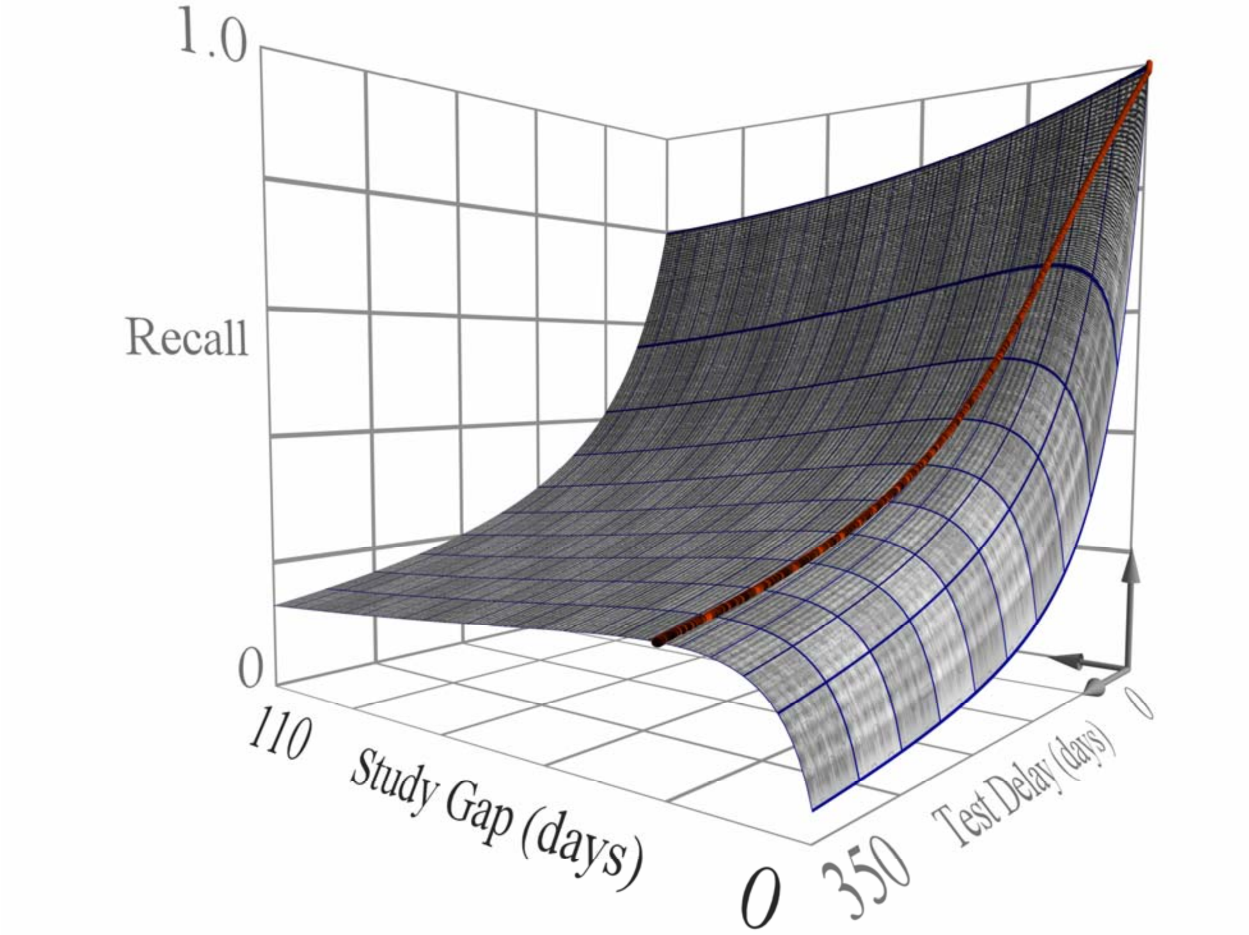

Figure 4