Research Article

Spacing Effects in Learning

A Temporal Ridgeline of Optimal Retention

Nicholas J. Cepeda,[1,2] Edward Vul,[2,3] Doug Rohrer,[4] John T. Wixted,[2] and Harold Pashler[2]

[1]York University; [2]University of California, San Diego; [3]Massachusetts Institute of Technology; and [4]University of South Florida

ABSTRACT—***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. In a study aimed at characterizing spacing effects over significant durations, more than 1,350 individuals were taught a set of facts and—after a gap of up to 3.5 months—given a review. A final test was administered at a further delay of up to 1 year. At any given test delay, an increase in the interstudy gap at first increased, and then gradually reduced, final test performance. The optimal gap increased as test delay increased. However, when measured as a proportion of test delay, the optimal gap declined from about 20 to 40% of a 1-week test delay to about 5 to 10% of a 1-year test delay. The interaction of gap and test delay implies that many educational practices are highly inefficient.***

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 fundamentally important if one wishes to temporally structure learning events in a rational manner. Taking the effects of the gap into account could yield important benefits if it turns out that these effects are large—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.

Address correspondence to Hal Pashler, Department of Psychology, University of California, San Diego, e-mail: hpashler@ucsd.edu.

Effects of the gap between exposures 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 (see Fig. 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, readers 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 reports. However, only about a dozen of these looked at RIs as long as 1 day, with just a handful examining RIs longer than 1 week. Although 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: On the basis of short-term studies, one cannot answer with confidence even basic questions about the timing of learning. For example, 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 RIs (Bahrick, Bahrick, Bahrick, & Bahrick, 1993), 4 subjects' acquisition and retention of foreign language vocabulary were examined over several years. 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 interstudy spacing to 56 days improved performance (see Fig. 2).

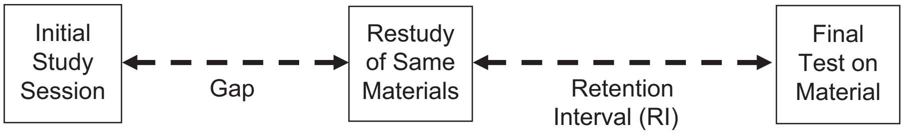

Fig. 1. Structure of a typical study of spacing effects on learning. Study episodes are separated by a varying gap, and the final study episode and test are separated by a fixed retention interval.

These studies might appear to suggest that over these long intervals, spacing effects may be monotonic, rather than showing an inverted-U shape, as found in shorter-term studies. However, 56 days is actually a relatively short proportion of the extremely long RIs Bahrick et al. used, and it is this ratio of gap to RI that is probably most critical, as the study reported here demonstrates.

Another issue in interpreting the studies by Bahrick et al. (1993) and Bahrick and Phelps (1987) is the fact that subjects were trained to a fixed performance criterion in each study session. Given the forgetting that takes place during the gap between study sessions, this procedure inevitably results in more relearning trials being provided in learning sessions separated by longer gaps. Although one might argue that in some cases students will wish to relearn to criterion, so that this procedure may be informative about the appropriate timing of such relearning sessions, this design feature makes it challenging to draw conclusions about the efficiency of study because it confounds total study time and spacing gap.

The goal of the present study was to examine the joint effects of gap and RI more systematically and over longer time intervals than has been done previously. We held constant the number of restudy trials in the second study session, which allowed 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 nonmonotonic relationship of retention to gap and, more generally, to reveal something about the shape of what we term the *retention surface*, that is, final test performance as a function of gap and RI.

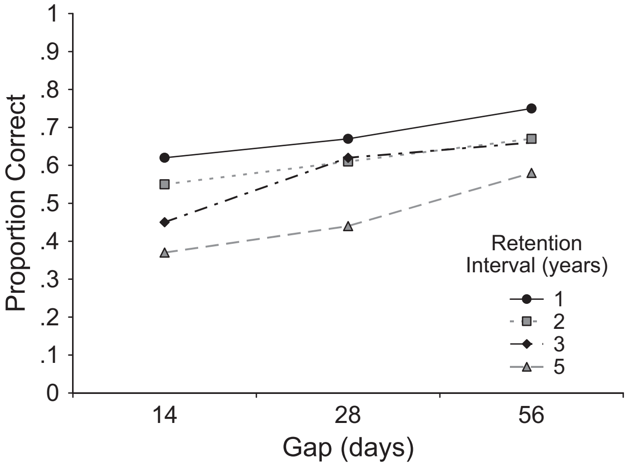

Fig. 2. Final test performance as a function of gap and retention interval in the study by Bahrick, Bahrick, Bahrick, and Bahrick (1993), which examined spacing over multiyear retention intervals.

This undertaking required running thousands of training and test sessions. Fortunately, the advent of Internet-based experimental testing panels has made 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 validity of Internet data collection has become increasingly clear in recent years. Although objections against this form of data collection are still occasionally raised, they receive little support from actual experience with the method.

PRELIMINARY DATA

In a preliminary laboratory-based study (Cepeda et al., in press) that provided a key benchmark for the present study, 150 subjects participated in three sessions over a period of up to 1 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 min to 6 months. All subjects then returned to the lab for a final memory test 6 months after their second learning session. The nonmonotonic pattern of results noted in short-term studies was indeed found: Recall success (for both facts and names) was best for a 1-month gap, being much worse for shorter gaps and slightly poorer for longer ones. If the optimal gap value should happen to increase linearly with the RI, then these results would imply that about a 15 to 20% ratio of gap to RI optimizes retention, but linearity cannot be assumed.

THE CURRENT STUDY

We now report the results of a more comprehensive set of learning episodes and tests involving 1,354 new subjects from our laboratory's Internet Memory Research panel, which was formed for long-term repeat testing. We suspect this may be the most systematic analysis of long-term spacing effects yet carried out. To properly characterize the interaction of gap and RI, we combined various gap and RI values, for a total of 26 different conditions. In the first learning session, subjects learned 32 facts to a criterion of one perfect recall for each fact. After the prescribed gap, a second learning session was completed. In this session, subjects were tested twice on each fact, and were shown the correct answer after they responded. After the prescribed RI, subjects were given two tests on each of the 32 facts, without feedback. The first was a recall test (e.g., Who invented snow golf?), and the second was a recognition test in which subjects tried to pick the correct answer from among five equally likely alternatives.

TABLE 1

Number of Subjects in Each Experimental Condition

Retention interval (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

Method

Subjects

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

We report data from subjects who completed all three sessions of the present study within the necessary time windows. Noncompletion rates increased at the longer delays, as one would expect in any multiyear study, but initial knowledge of the facts did not differ reliably between subjects who completed all three sessions and those who did not complete the final test. There were also no detectable differences between these two groups in age, gender, number of obscure facts known before beginning the study, or a wide array of background and demographic characteristics. The mean age of the subjects who completed the study was 34 years (SD = 11, range = 18–72), and 72% were female.

Stimuli and Materials

The 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 five or six letters. As Table 1 shows, the study included a range of gaps (interval between the first and second learning sessions) and RIs (interval between the second learning session and the final test).

Design and Procedure

There were 26 gap-by-RI combinations, and each subject was randomly assigned to one of these. The number of gaps for each RI varied: six gaps for each of two RIs and seven gaps for each of the other two RIs. RIs and gaps were chosen such that there were five gaps in common to all the RIs and each RI 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 the conditions requiring longer time intervals between sessions, in order to compensate for the anticipated greater noncompletion 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; subjects were encouraged to guess if they were not confident of the answer, and then the correct answer was provided as feedback. The first presentation of each fact allowed us to identify and remove from analysis any items known to a given 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. In each cycle, the items were presented in a new random order. Subjects answered between 61 and 96 questions in the course of the first session before they reached the criterion.

Subjects were advised by e-mail when it was time for them to perform the second session. When the 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 0.00256 days). In the second learning session, the same entire list of questions was run through twice, each time in a different random order. Each item was followed by a presentation of the correct answer. Subjects could take as long as they wished to answer, or leave the item blank. Regardless of a subject's response, the correct answer was displayed for 4 s, and the next question appeared after approximately 1 s.

During the final session, subjects were given two tests, each covering all 32 facts. No feedback was provided in this phase. The first test 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 in a separate pilot study with subjects 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 RIs, for the subjects who completed all

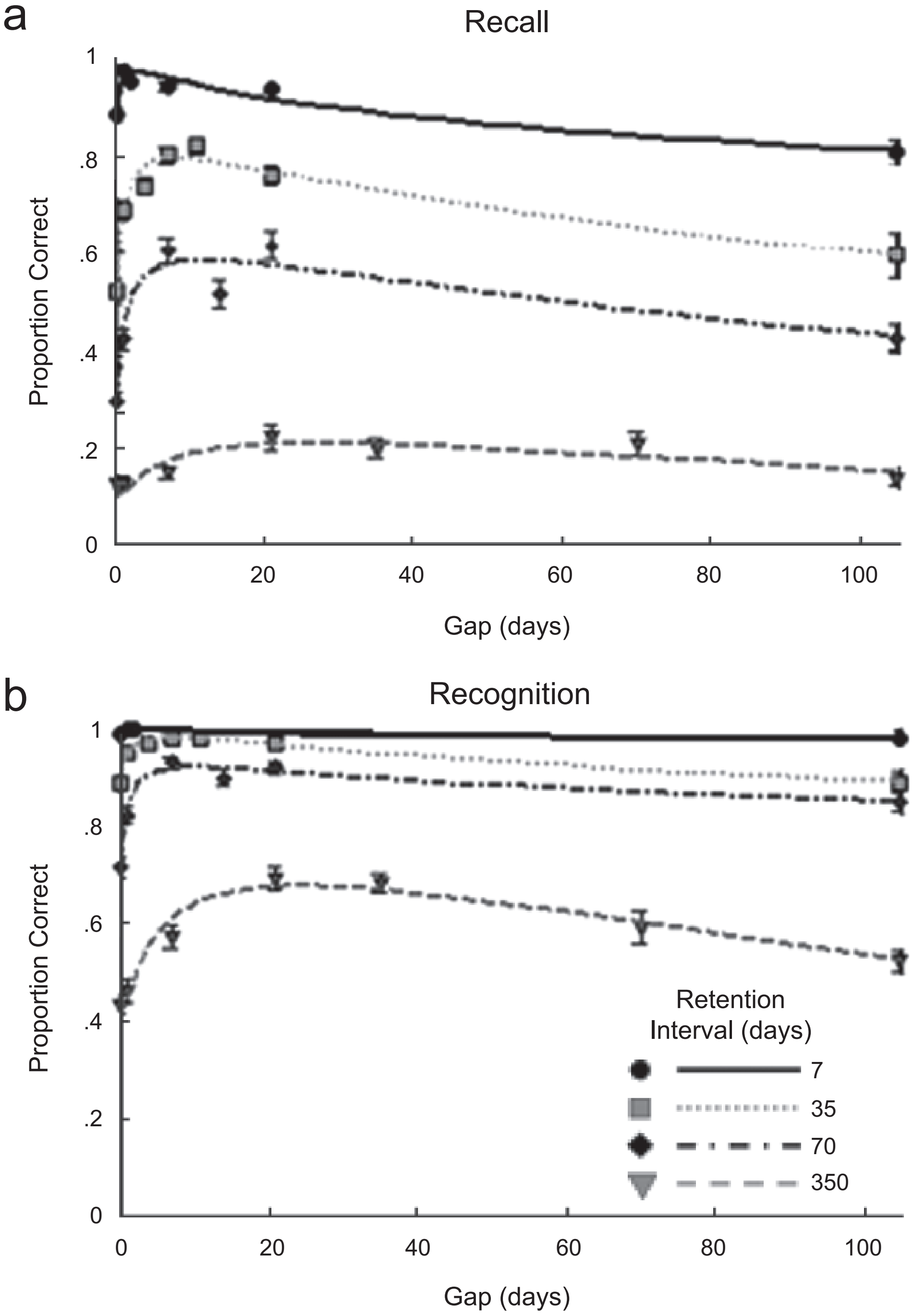

Fig. 3. Performance on the final (a) recall and (b) recognition tests as a function of gap, for each of the four retention intervals. The plotted points show mean accuracy ± 1 *SEM*. The lines correspond to cubic spline fits to the data, with fixed points at gaps of 0 and 105 days.

the phases of the study. For each RI, final performance initially rose with increasing gap and then fell as gap was increased further. The effects of gap were very large in magnitude: For a fixed amount of study time, the optimal gap, as compared with a zero-day gap, provided a 64% increase in final recall, $d = 1.1$, and a 26% increase in final recognition, $d = 1.5$ (in this article, d values refer to the comparison of the zero-day and optimal gaps). 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 and 1, 7, 7, and 21 days, respectively, for recognition. We were able to obtain more precise estimates of the optimal gaps by interpolating our data with cubic splines (see Fig. 3); for recall, these interpolated gaps were approximately 3, 8, 12, and 27 days (corresponding to 43%, 23%, 17%, and 8% of the RIs, respectively), and for recognition, the interpolated gaps were approximately 1.6, 7, 10, and 25 days (24%, 19%, 14%, and 7% of the RIs). All these findings are in generally good agreement with our findings in the lab-based benchmark study. The value of the estimated optimal gap rose as RI increased, and the optimal gaps departed noticeably from the fixed proportion of the RI suggested by some earlier researchers on the basis of much shorter-term studies (Crowder, 1976; Murray, 1983).

The 7-, 35-, 70-, and 350-day RIs yielded 10, 59, 111, and 77% improvement in recall for the optimal gap, as compared with the zero-day gap. The improvement was reliable in each case, $t(124) = 6.5, p_{\text{rep}} = .99, d = 1.3$; $t(111) = 8.9, p_{\text{rep}} = .99, d = 0.6$; $t(102) = 8.6, p_{\text{rep}} = .99, d = 1.7$; $t(68) = 3.9, p_{\text{rep}} = .99, d = 0.9$, respectively. The corresponding improvements in recognition were 1, 10, 31, and 60%. Again, the improvement was reliable in each case, $t(124) = 2.3$, $p_{\text{rep}} = .99$, $d = 0.7$; $t(136) = 7.5, p_{\text{rep}} = .99, d = 1.5$; $t(112) = 8.7, p_{\text{rep}} = .99, d = 1.7$; $t(68) = 7.9, p_{\text{rep}} = .99, d = 2.1$.

DISCUSSION

The results presented here document the existence of very large and nonmonotonic spacing effects that unfold over very long periods of time, when study time is 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 RI. 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 suggested 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 Fig. 3a), as shown by the direction of the red ridgeline in Figure 4. Fourth, as RI increases, the ratio of optimal gap to RI should decline. In Figure 4, for example, the optimal gap for the 350-day RI is 23 days, which is just 7% of the RI.

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

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

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

$$A = p + (1 - p)e^{-ag},$$

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, which means 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 Fig. 4 has the following parameter values: $b = 0.011, p = 0.760$, $a = 0.017$, $c = 0.092$, and $d = 3.453$.) A number of other functions can also provide quite decent fits to the data, with various trade-offs 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 the surface in Figure 4, seen over such long intervals, may help in constraining theories concerning the mechanisms of the spacing effect.[1] 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), do not seem to fit well with the 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 days and weeks (the optimal gap was several weeks for our longer RIs). 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. Such theories suggest that spacing will benefit learning when the first learning episode has been forgotten; thus, longer study gaps should always produce better final retention, and there should not be an optimal gap. Other distributed-practice theories, such as encoding variability (Glenberg, 1979) and study-phase re-

[1] Our results also probably help explain why, as noted earlier, Bahrick et al. (1993) did not observe nonmonotonic effects of gap: The largest ratio of gap to RI in their study was only 0.15 (56 days/365 days), which (in light of data from the present study) might well be insufficient to show the declining segment of the gap effect.

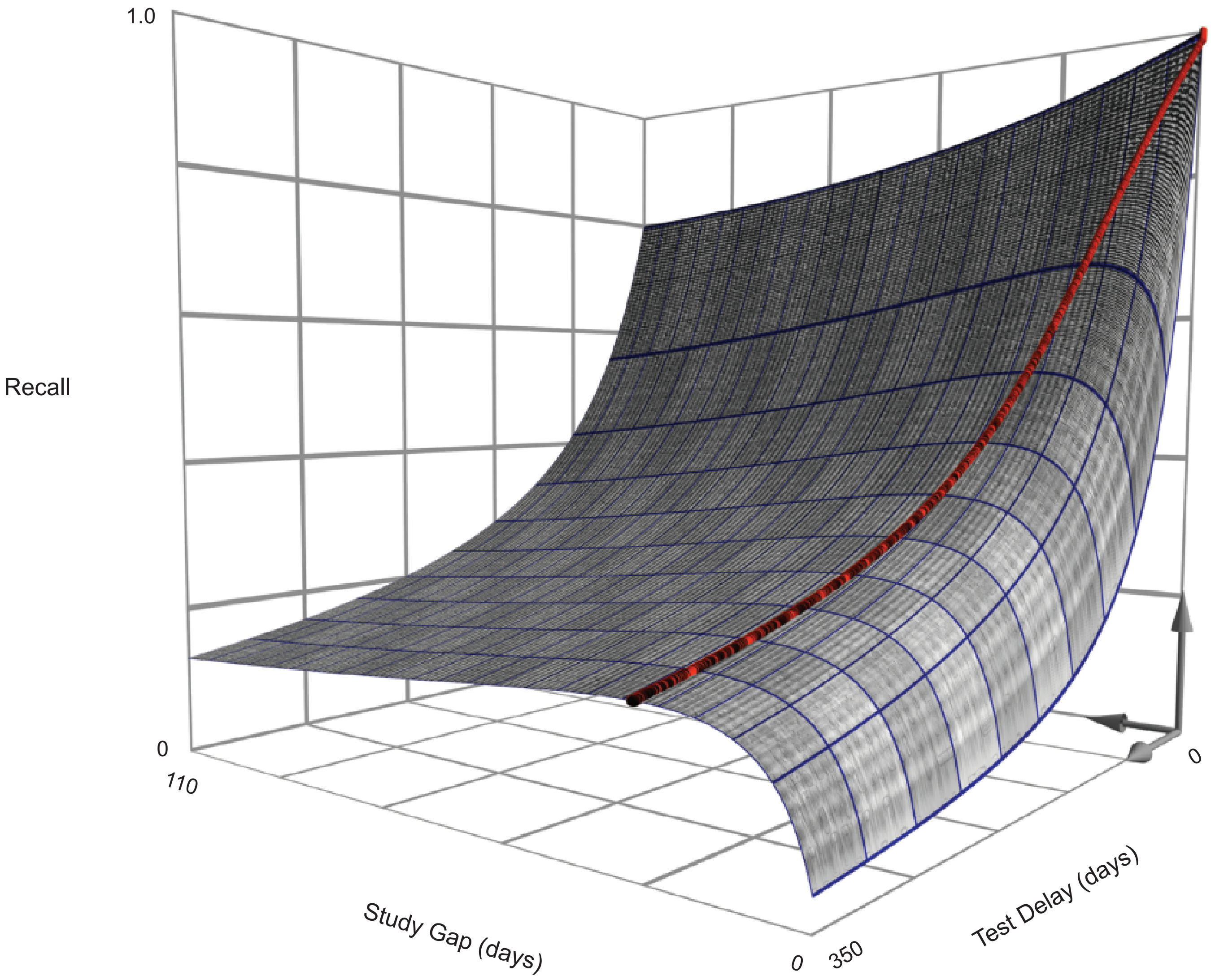

Fig. 4. A functional approximation of recall on the final test (as a proportion), plotted as a function of gap and test delay (i.e., retention interval). The red ridgeline comprises the points representing the optimal performance for each test delay. The forgetting function for each gap is a power function.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

trieval (Murray, 1983), are potentially consistent with the basic results shown 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 performance, but also performance in the second learning session (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 the ratio of optimal gap to RI as a function of increasing RI. Whether or not this conclusion stands, it seems likely that the present data provide significant new constraints on theorizing about memory and spacing effects over meaningful time intervals.

Educational Implications

The present results show that the timing of learning sessions can have powerful effects on retention when study time is equated, and these effects, as in our benchmark study, seem far larger than those typically seen in studies using short-term spacing (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 rather depends dramatically on the RI (a point that was evident in the short-term studies, such as that by Glenberg, 1976, and is now shown to extend to far greater time intervals). To put it simply, if you want to know the optimal distribution of your study time, you need to decide how long you wish to remember something.

Although the interactive effects of gap and RI pose 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 the time investment—potentially doubling the amount ultimately remembered compared with a less temporally distributed study schedule, with study time equated. Although this advice is in agreement with the earlier work of Bahrick (e.g., Bahrick et al., 1993), it is at odds with many conventional educational practices—for example, study of a single topic being confined within a given week of a course. The current results indicate that 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 (as some researchers have long surmised; see, e.g., Bahrick, 2005; Dempster, 1988; and Schmidt & Bjork, 1992). It is also of interest to note that although 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 short a gap, as evidenced by the fact that as the gap increases, accuracy increases steeply and then declines much more gradually (see Fig. 3). In light of the present results, it appears no longer premature for psychologists to begin to offer some rough practical guidelines to people who wish to use study time in the most efficient way possible to promote long-term retention.

Acknowledgments—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, principal investigator; Grant SBE-0542013, G.W. Cottrell, principal investigator). The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education. We thank Harry Bahrick and an anonymous reviewer for helpful comments on this manuscript.

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., Bahrick, L.E., Bahrick, A.S., & Bahrick, P.E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

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

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. (in press). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology.*

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: Erlbaum.

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 & 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., & 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.

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.

(Received 10/3/07; Revision accepted 3/28/08)

APPENDIX: VALIDITY OF INTERNET-BASED MEMORY TESTING

Internet testing has become common in the behavioral sciences over the past several years, and standards based on early experiences with this method have now been developed. Our Internet testing procedures followed the standards recommended by Reips (2002).

The validity of Internet testing has been well supported in recent reviews (Gosling, Vazire, Srivastava, & John, 2004), and excellent correspondence between results obtained with Internet samples and results obtained in the laboratory has been reported (e.g., Birnbaum, 1999; Krantz & Dalal, 2000; McGraw, Tew, & Williams, 2000; Reips, 2002). These reports track our own experience in conducting memory studies in the lab and on the Web. This correspondence between Web-based and lab-based data was also demonstrated by the Internet-based results and the laboratory-based benchmark data discussed in the main text. It is our 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 against Internet testing are sometimes raised, 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 scores obtained with Internet samples and laboratory samples does not suggest any important differences. For example, for conditions with roughly the same gaps and retention intervals (RIs), average performance on the final test was 41% in the Internet study reported in this article and 45% in the benchmark study.

Another potential concern sometimes raised is the possibility of "cheating" (i.e., writing down answers). Note that because of the randomized between-subjects design we used in this study, even if there were some small incidence of cheating or more severe distraction than occurs in the lab, these elements would have merely introduced noise, thus dampening the effects of the temporal variables—effects that (as we have discussed) 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 study and 2.1% for comparable gap-by-RI cells 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, although it is understandable for researchers to view Web-collected data with initial caution, actual experience with Internet-based methods provides little reason to believe that such data are any less credible than lab-collected data.