Optimizing Distributed Practice

Theoretical Analysis and Practical Implications

Nicholas J. Cepeda,[1,2] Noriko Coburn,[2] Doug Rohrer,[3] John T. Wixted,[2]
Michael C. Mozer,[4] and Harold Pashler[2]

[1]York University, Toronto, ON, Canada
[2]University of California, San Diego, CA
[3]University of South Florida, FL
[4]University of Colorado, Boulder, CO

Abstract. More than a century of research shows that increasing the gap between study episodes using the same material can enhance retention, yet little is known about how this so-called *distributed practice effect* unfolds over nontrivial periods. In two three-session laboratory studies, we examined the effects of gap on retention of foreign vocabulary, facts, and names of visual objects, with test delays up to 6 months. An optimal gap improved final recall by up to 150%. Both studies demonstrated nonmonotonic gap effects: Increases in gap caused test accuracy to initially sharply increase and then gradually decline. These results provide new constraints on theories of spacing and confirm the importance of cumulative reviews to promote retention over meaningful time periods.

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

Introduction

An increased temporal lag between study episodes often enhances performance on a later memory test. This finding is generally referred to as the "spacing effect", "lag effect", or "distributed practice effect" (for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1989; Dempster & Perkins, 1993; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003; Moss, 1995). The distributed practice effect is a well-known finding in experimental psychology, having been the subject of hundreds of research studies (beginning with Ebbinghaus, 1885/1964; Jost, 1897). Despite the sheer volume of research, a fundamental understanding of the distributed practice effect is lacking; many qualitative theories have been proposed, but no consensus has emerged. Furthermore, although distributed practice has long been seen as a promising avenue to improve educational effectiveness, research in this area has had little effect on educational practice (Dempster, 1988, 1989; Pashler, Rohrer, Cepeda, & Carpenter, 2007).

Presumably for reasons of convenience, most distributed practice studies have used brief spacing gaps and brief retention intervals, usually on the order of seconds or minutes. Few data speak to retention overnight, much less over weeks or months. Therefore, there is little basis for advice about how to maximize retention in real-world contexts. To begin to fill this notable hole in the literature, we present two new experiments that examine how the duration of the spacing gap affected the size of the distributed practice effect when the retention interval was educationally meaningful.

Distributed Practice: Basic Phenomena

The typical distributed practice study – including the studies described below – requires subjects to study the same material in each of the two learning episodes separated by an interstudy gap (henceforth, *gap*). The interval between the second learning episode and the final test is the test delay. In most studies, the test delay is held constant, so that the effects of gap can be examined in isolation from test delay effects.

A recent literature review (Cepeda et al., 2006) found just 14 studies that provided comparisons of very short (<3 h) and long (1 day or more) gaps with test delays of 1 day or more (Bahrick, 1979; Bahrick & Phelps, 1987; Bloom & Shuell, 1981; Childers & Tomasello, 2002; Fishman, Keller, & Atkinson, 1968; Glenberg & Lehmann, 1980; Gordon, 1925; Harzem, Lee, & Miles, 1976; Keppel, 1964; Robinson, 1921; Rose, 1992; Shuell, 1981; Watts & Chatfield, 1976; Welborn, 1933). In each study, a one-or-more day gap was superior to a very short gap. Thus, the extant data suggest that a gap of <1 day is reliably less effective than a gap of at least 1 day, given a test delay of 1 day or more.

Is a 1-day gap sufficient to produce most or even all the distributed practice benefits? To answer this question, we reviewed studies that used multiple gaps of 1 day or more, with a fixed test delay of at least 1 day. Thirteen studies satisfy these criteria (Ausubel, 1966; Bahrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Phelps, 1987; Burtt & Dobell, 1925; Childers & Tomasello, 2002; Edwards, 1917; Glenberg & Lehmann, 1980; Simon, 1979;

Experimental Psychology 2009; Vol. 56(4):236–246
DOI: 10.1027/1618-3169.56.4.236

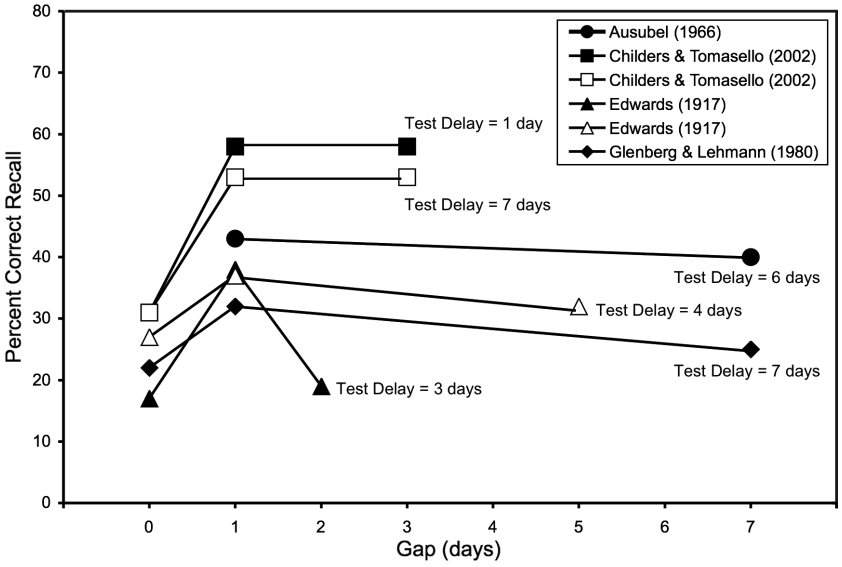

Figure 1. Percentage of items recalled during the final retention test, for prior unconfounded experiments. A 1-day gap produced optimal retention at the final test.

Spitzer, 1939; Strong, 1916, 1973; Welborn, 1933). We found that many of these 13 studies had undesirable methodological features. For instance, several studies trained subjects to a performance criterion on Session 2, and the presumed increase in total study time after longer gaps confounds these studies. As an example of this problem, Bahrick et al. (1993) reported that subjects required twice as many trials in the second study session in order to achieve criterion, as gap increased from 14 to 56 days. Also problematic, Welborn (1933) omitted feedback from Session 2, implying that the second session probably provided no opportunity for learning those items that are not learned during Session 1 (Pashler, Cepeda, Wixted, & Rohrer, 2005). Once these problematic studies were excluded, just four studies remain (Figure 1; Ausubel, 1966; Childers & Tomasello, 2002; Edwards, 1917; Glenberg & Lehmann, 1980). These studies suggest that a gap of roughly 1 day is optimal, but they hardly demonstrate this claim with any certainty, especially given the restricted set of test delays used.

The possibility that test accuracy might follow an inverted U-function of gap has been suggested by Balota, Duchek, and Paullin (1989), Glenberg (1976), Glenberg and Lehmann (1980), and Peterson, Wampler, Kirkpatrick, and Saltzman (1963). In this figure there are several possibilities. First, a fixed gap (e.g., 1 day) might be optimal, regardless of the test delay, which means that a gap less than or greater than 1 day would produce less than optimal test scores. Indeed, the studies shown in Figure 1 at the first glance suggest that a 1-day gap is always optimal. Second, optimal gap might be a fixed proportion of test delay (e.g., 100% of the test delay; Crowder, 1976; Murray, 1983), although a solid empirical or theoretical case for a ratio rule has not been offered.[1] Third, optimal gap might vary with test delay in some other way that would not conform to a ratio rule. For example, the optimal gap might increase as a function of test delay and yet be a declining proportion of test delay.

Theoretical Constraints

Because no one has quantitatively characterized the nature of distributed practice functions over time intervals much beyond a day, existing theories of distributed practice may not have much bearing on the phenomenon as it arises over a much longer time period. Indeed, some existing distributed practice theories were formulated in ways that seem hard to apply to gaps longer than a few minutes. For example, many theories (e.g., all-or-none theory; Bower, 1961) focus on the presence or the absence of items in working memory. If distributed practice benefits retention at gaps far exceeding the amount of time an item remains in working memory, then such theorizing must be incomplete at best. Including gaps of at least 1 day insures that the range includes at least one night of sleep, which may play a significant role in memory retention (Peigneux, Laureys, Delbeuck, & Maquet, 2001).

[1] Crowder (1976), based on the Atkinson and Shiffrin (1968) model of memory, stated that "the optimal [gap] is determined by the delay between the second presentation and the testing. If this testing delay is short, then massed repetition is favored but if this delay is longer then more distributed schedules of repetition are favored" (p. 308). Murray (1983), based on Glenberg (1976, 1979), stated that "spacing facilitates recall only when the retention interval is long in proportion to the [gap], and that recall decreases with [increased gap] if the [gap is] longer than the retention interval" (pp. 5–6).

Overview of Experiments

The studies reported here assessed the effects of gap duration on subsequent test scores with moderately long gaps and test delays. In Experiment 1, the test delay was 10 days, and gaps ranged from 5 min to 14 days. These values are roughly equal to those used in the four studies as shown in Figure 1; thus, Experiment 1 allows us to compare our results with prior findings and expands the sparse literature using meaningful test delays. Experiment 2 used a 6-month test delay and gaps ranging from 20 min to 168 days. Experiments 1 and 2 are the first unconfounded examinations of paired associate learning in adults, using day-or-longer test delays. By comparing the results of these studies, we can tentatively support or refute the claim that optimal gap varies with test delay, as suggested previously (Crowder, 1976; Murray, 1983).

Experiment 1

The first study examined how retention is affected as gap is increased from 5 min to 14 days, for a test delay of 10 days. Subjects learned Swahili-English word pairs; the Swahili language was selected because English speakers can readily articulate Swahili words even though the language is entirely unfamiliar to most students at the University of California, San Diego. (When asked, no subjects in our sample reported prior exposure to Swahili.)

Method

Subjects

A total of 215 undergraduate students from the University of California, San Diego, enrolled in a three-session study. Those who finished all three sessions (n = 182) received course credit and US $6.00 payment. There were 31, 31, 30, 29, 29, and 32 subjects who yielded usable data in the 0-, 1-, 2-, 4-, 7-, and 14-day gap conditions, respectively.

Materials

Subjects learned the (single-word, 3–10 letters) English translations for forty 4–11 letter Swahili words.

Design

Subjects were randomly assigned to one of six conditions (0-, 1-, 2-, 4-, 7-, or 14-day gap), and for the 0-day condition, the gap was ~5 min.

Procedure

Subjects completed two learning sessions and one test session. They were trained and tested individually on a computer located in a sound-attenuated chamber. Figure 2 shows the overall procedure for each experimental session. The first session began with instructions stating "You will be learning words from a foreign language. First you will see the foreign word and its English translation. Try to remember each correct English translation. You will be tested until you correctly translate each foreign word two times. The correct translation will appear after you make your response." Immediately afterwards, subjects saw all 40 Swahili-English word pairs, presented one at a time in a random order, for 7 s each, with each Swahili word appearing directly above its English translation. Then, subjects began test-with-feedback trials in which they repeatedly cycled through the list of Swahili words and attempted to recall the English equivalent for each Swahili word. Subjects were prompted to type the English equivalent immediately after seeing each Swahili word. Subjects could take as long as needed to type their response. Immediately after a response was made, the computer sounded a tone indicating a correct or incorrect response, and both the Swahili word and its English equivalent appeared on the screen for 5 s (regardless of whether the subject had responded correctly). After two correct responses were made for a given word (although not necessarily on consecutive list presentations), the word was not presented again. Subjects continued to cycle through the list (in a new random order each time) until there were no items left.

Depending on the gap, each subject returned for the second learning session between 5 min and 14 days later. The second learning session consisted of two cycles through the list of Swahili words, with each cycle including a test-with-feedback trial for each word. Again, unlimited response time was allowed. Auditory feedback followed immediately after each response, and visual feedback (the correct answer) was displayed for 5 s following each response. The entire list of 40 word pairs was tested with feedback, two times, in a different random order each time (different for each subject). (Subjects were not taught to criterion in the second learning session, as they were in the first, because that would have confounded the gap and the number of trials required during the second session, as explained in the Introduction section.)

Subjects returned for the test session 10 days after the second session. (If the 10th day fell on a weekend, the test was shifted to the nearest weekday.) Subjects were again instructed to type the English translation for each Swahili word. Unlike in the learning sessions, feedback was not provided. The Swahili words appeared in a random order, which was different for each subject, and each word was tested once.

Results and Discussion

Figure 3 shows the performance on the first test of Session 2 and the Session 3 test (administered 10 days after Session 2). The first test of Session 2 measured retention after a single exposure period, and these data therefore show a traditional forgetting function. For the final test, which reflects the benefits of spacing, a 1-day gap optimized recall.

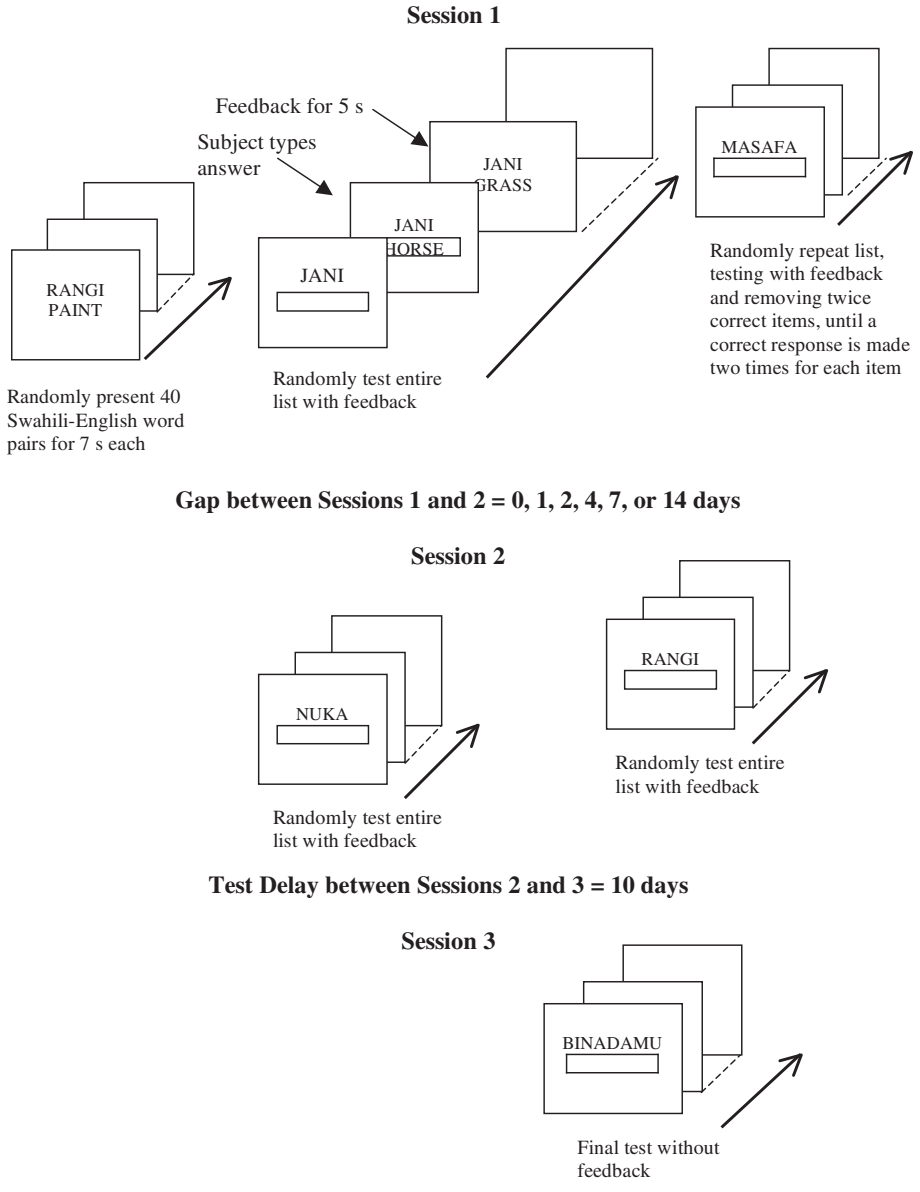

Figure 2. Experiment 1 procedure.

Moreover, varying gap had a large effect: Recall improved by 34% as gap increased from 0 to 1 day. Increases in gap beyond a single day produced a small but relatively steady decline in final-test scores, with recall accuracy decreasing just 11% as gap increased from 1 to 14 days.

These distributed practice effects were analyzed in several different ways. First, effect sizes were computed for each adjacent pair of gaps (Table 1). These effect sizes show the large benefit of increasing gap from 0 to 1 day. Second, a one-way analysis of variance (ANOVA) was conducted, using final-test recall as a dependent variable and gap as an independent variable. There was a main effect of gap, $F(5, 176) = 3.7$, $p < .005$. Third, Tukey Honestly Significant Difference (HSD) tests show that the 0-day gap produced significantly worse recall than the 1-, 2-, 4-, and 7-day gaps; no other pair-wise comparisons were significant.

The results show generally good agreement with previous confound-free studies that used similar gaps and test delays, as shown in Figure 1 (i.e., Ausubel, 1966; Childers & Tomasello, 2002; Edwards, 1917; Glenberg & Lehmann, 1980). It appears that the nonmonotonic relationship between gap and memory retention generalizes well from the text recall (Ausubel), object recall (Childers & Tomasello), fact recall (Edwards), and free recall of word lists (Glenberg & Lehmann) to associative memory for foreign language vocabulary. However, because these four studies and Experiment 1 used approximately equal test

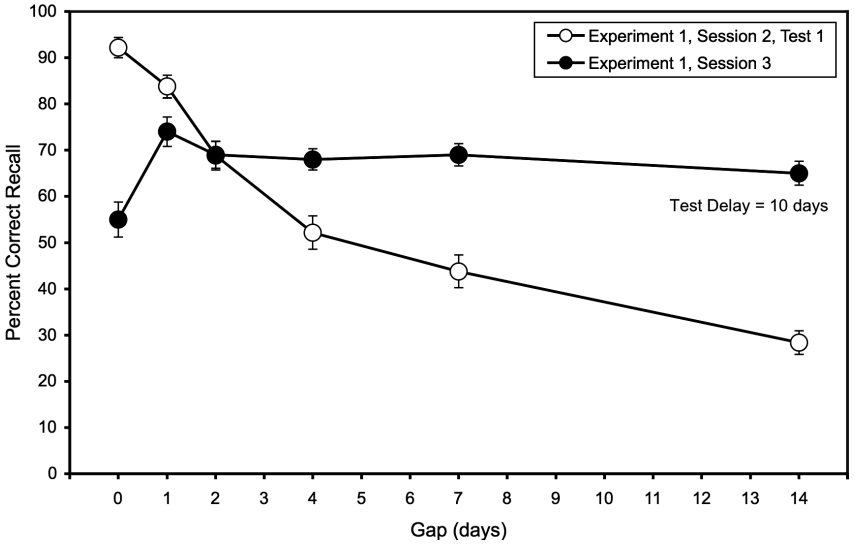

Figure 3. Percentage of items recalled during the first test of Session 2 and the final retention test, for Experiment 1. Bars represent one *SEM*. A 1-day gap produced optimal retention at the final test.

Table 1. Effect size (Cohen's *d*) and change in percent correct (PC) between different gaps, for Experiment 1. Gap shows days between learning sessions

Gap			
Short	Long	*d*	PC
0	1	1.03	18.9
1	2	−0.28	−5.0
2	4	−0.02	−0.4
4	7	0.06	1.0
7	14	−0.22	−4.0
1	14	−0.46	−8.4

delays, the possibility remains that a much longer test delay would yield an optimal gap other than 1 day. This possibility was examined in Experiment 2.

Experiment 2

The second study used a much longer test delay (6 months) than Experiment 1. Because pilot data suggested that Swahili-English word pairs (which were used in Experiment 1) would produce floor effects after a 6-month test delay, we chose material that was shown to produce lesser rates of forgetting. The material was again educationally relevant: Not-well-known facts and names of unfamiliar visually presented objects. The two study sessions were separated by gaps ranging from 20 min to 6 months, with the final-test given 6 months after the second study session.

Method

Subjects

A total of 233 undergraduates from the University of California, San Diego, began the study. Those who finished all three sessions received US $30 payment. Data from 72 subjects were discarded (37 because they failed to complete all three sessions, 34 because they did not complete Session 2 or 3 within the allotted time frame, and 1 because he began working in our laboratory and was no longer considered blind to the purpose of the study). Table 2 shows fewer subjects in the 6-month gap condition, partly due to the increased difficulty maintaining contact with these subjects; otherwise, dropout rates did not vary across conditions.[2] Of the 161 subjects included in the analyses, 66% were female, and the mean age was 19.6 years (*SD* = 2.4). None of the Experiment 2 subjects had participated in Experiment 1.

Materials

For Part A, a list of 23 not-well-known facts was assembled. Each fact was presented as a question and then an answer.

[2] Subjects in the 6-month gap condition were equivalent to other subjects on a wide range of demographic measures. Even if 6-month gap subjects' memory performance was better than their cohorts' memory performance (and our analyses suggest it wasn't), further analysis of covariance removed the effects of differential memory ability across subjects and showed the same effects. Our conclusions do not depend on the performance of the 6-month gap group. The data from this group are qualitatively and quantitatively consistent with the 3-month gap data and the literature more broadly.

Table 2. Actual gaps and test delays for each experimental group, for Experiment 2

Gap group	No. of subjects	Mean gap (range in days)	Mean test delay (range in days)
0 day	28	20[a] (none)	168.6 (161–179)
1 day	34	1 (none)	171.0 (160–181)
1 week	29	6.9 (6–8)	165.5 (159–176)
1 month	23	28.5 (23–34)	168.1 (160–180)
3 months	31	83.0 (77–90)	166.1 (158–176)
6 months	16	169 (162–175)	167.8 (156–181)

[a]Data in minutes.

For example, the fact "Rudyard Kipling invented snow golf" was presented as "Who invented snow golf?" and "Rudyard Kipling". For Part B, a set of 23 photographs of not-well-known objects was assembled. For example, objects included a "Lockheed Electra" airplane. Each photo was associated with a question and a fact: For example, "Name this model, in which Amelia Earhart made her ill fated last flight" and "Amelia Earhart made her ill fated last flight in this model of Lockheed Electra". A clipboard, pen, and paper with prenumbered answer blanks were provided during testing.

Design

Subjects were randomly assigned to one of six conditions (0-, 1-, 7-, 28-, 84-, or 168-day gap), and for the 0-day condition, the gap was ~20 min.

Procedure

The experiment was conducted in a simulated classroom setting in a windowless room. A computer-controlled liquid crystal display projector displayed the stimuli on one wall of the room, and prerecorded audio instructions and audio stimuli were presented (simulating the "teacher") through speakers placed in front of the room. A computer program controlled the presentation of visual and auditory stimuli. An experimenter initiated each section of the experiment, answered questions about the instructions, and monitored subjects' compliance with the instructions. Subjects were tested in groups of 1–6.

Subjects were informed that we were examining changes in learning over time, that 23 items would be presented, that items might change across sessions, and that there would be a series of tests, with feedback, to help them learn the items. They were asked to write each answer in the appropriate answer box and were asked not to change the answer after feedback began. Subjects were informed that there was no penalty for incorrect guesses or partial answers. During each session, all obscure facts (Part A) preceded all visual objects (Part B).

In Session 1, the instructions were followed by a pretest, one initial exposure to each of the 23 items, and then three blocks of 23 test-with-feedback trials. In each block of 23-item presentations, a new random order was used; this random order was constant across subjects. For the pretest, each fact was visually presented as a question (13 s) as the "teacher" read the fact. Then this answer sheet was collected by the experimenter. Immediately afterward, each of the 23 items appeared on the screen in statement form (13 s) as the "teacher" read the statement. This was followed immediately by the three blocks of test-with-feedback trials. For each of these trials, the subjects first saw either a question (Part A) or a photo (Part B) for 13 s, during which time the question or the associated fact was spoken by the "teacher". During this interval, subjects attempted to write their answers in a space provided on their answer sheets. Immediately afterwards, the correct answer appeared (5 s) and was spoken by the "teacher". After each of the three blocks of test-with-feedback trials, the answer sheet was collected by the experimenter. Session 2, by contrast, included no pretest or learning trial, and subjects completed just two blocks of test-with-feedback trials. During Session 3, items were tested without feedback, first using a recall test and then using a multiple-choice recognition test with four possible answers. Pilot testing confirmed that the options in the multiple-choice test were about equally likely to be chosen by subjects with no previous knowledge of the fact or the object.

Results and Discussion

The range of actual gaps and test delays and average gaps and test delays are shown in Table 2; these differed slightly from the nominal gaps and test delays listed in our design because of our inability to schedule some subjects' second or third session on precisely the desired day.

Each response was scored by "blind" research assistants who were given a set of predetermined acceptable answers. Each item was assigned a score for correct answer, incorrect answer, or nonresponse (no answer). In general, misspellings were allowed (such as "Elektra" instead of "Electra"), and partial answers were considered correct when distinctive parts of the complete answers were given (e.g., "Ranger" for "USS Ranger"). Before the final data analysis, a single research assistant rechecked all difficult-to-code items, in order to confirm that all coders used identical scoring criteria across all subjects. As well, research assistants checked each others' work and discussed how to code difficult answers with each other and with the principal investigator (N.J.C.). All coding was done blind to experimental condition.

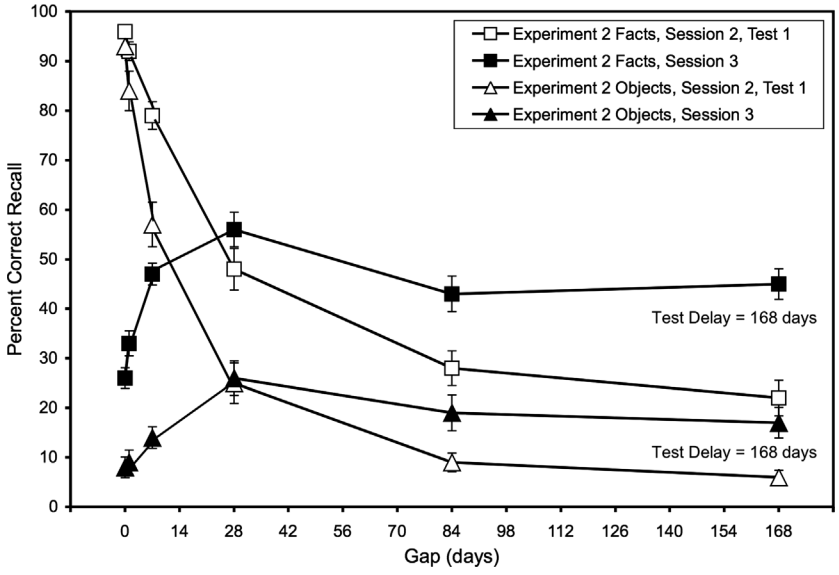

Figure 4. Percentage of items recalled during the first test of Session 2 and the final retention test, for Experiment 2. Bars represent one *SEM*. A 1-month gap produced optimal retention at the final test.

For each subject, items that were answered correctly during the pretest were excluded from analysis of their data, leading to the exclusion of <1% of items, on average. Performance on the first test of Session 1 showed no main effect or interaction involving gap. Facts were easier to learn than pictures, $F(1, 155) = 254.9$, $p < .001$. First-test accuracy ranged from 75% to 82% by gap for facts and from 45% to 64% by gap for objects. Likewise, performance on the third and final test of Session 1 showed no main effect or interaction involving gap, although facts showed slightly greater learning than pictures (94% vs. 90%), $F(1, 155) = 10.3$, $p < .005$. (The percentage of items learned during Session 1 was probably higher than this, because additional learning occurred from the final test. Figure 4 (Session 2, Test 1, 0-day gap) shows 96% and 93% accuracy for facts and objects, even after a 20 min delay.)

Figure 4 shows performance on the first test of Session 2 and the Session 3 test (6 months after Session 2). As in Experiment 1, performance on the first test of Session 2 exhibited a typical forgetting function. In contrast to the results of Experiment 1, final-test recall performance was optimized by a gap of *28 days* rather than just 1 day. In fact, the 28-day gap produced 151% greater retention than the 0-day gap, whereas the 1-day gap produced only an 18% improvement over the 0-day gap. Increasing gap from 28 to 168 days produced a relatively modest decline in retention of only 23%.

The effects of gap on recall were analyzed in several different ways. First, effect sizes were computed for each adjacent pair of gaps (Table 3). These effect sizes show the large benefit of increasing gap from 0 to 28 days. Second, a mixed-model ANOVA was conducted using final-test recall as a dependent variable, gap as a between-subjects factor, and type of material (facts or objects) as a within-subjects factor. There were main effects of gap, $F(5, 155) = 8.3$, $p < .001$, and material, $F(1, 155) = 502.2$, $p < .001$, and an interaction between gap and material, $F(5, 155) = 4.6$, $p < .005$. The interaction between gap and material likely reflects the different degrees of improvement, relative to baseline, for fact versus visual object materials; there are no obvious qualitative differences in the results. Third, Tukey HSD tests show that the 0-day gap produced significantly worse recall than all gaps longer than 1 day. The 1-day gap produced significantly worse recall than the 28-day gap. No other pair-wise comparisons were significant. This suggests that the 28-day gap was optimal and supports a claim that final-test recall gradually declines with too-long gaps. Quite dramatically, this demonstrates that a 1-day gap is not always optimal, since 0- and 1-day gaps were not significantly different, and recall was significantly worse for 1-day versus 28-day gaps.

Table 3. Effect size (Cohen's *d*) and change in percent correct (PC) between different gaps, for Experiment 2 recall data. Gap shows days between learning sessions

Gap			
Short	Long	*d*	PC
0	1	0.23	3.0
1	7	0.77	9.8
7	28	0.80	12.6
28	84	−0.57	−11.3
84	168	0.08	1.6
28	168	−0.25	−4.8
0	28	1.56	25.5
28	168	0.51	−9.8

For the multiple-choice recognition test, a mixed-model ANOVA was conducted using final-test recognition as a dependent variable, gap as a between-subjects factor, and

type of material (facts or objects) as a within-subjects factor. There was a main effect of gap, $F(5, 155) = 4.9$, $p < .001$. Recognition test performance at 0-, 1-, 7-, 28-, 84-, and 168-day gaps was 91 (9.1), 95 (5.1), 97 (3.2), 98 (2.4), 95 (5.2), and 96 (7.2) percent correct (*SD*), respectively, mirroring the recall test results.

General Discussion

Two experiments examined how the gap separating the two study episodes affected performance on a subsequent test given as much as 6 months later. Three primary novel findings are reported: First, spacing benefits were seen with test delays longer than 1 week (Figures 3 and 4), using a nonconfounded design. Second, gap had nonmonotonic effects on final recall even with test delays longer than a week; accuracy first increased and then decreased as gap increased. Third, for sufficiently long test delays, the optimal gap exceeds 1 day, whereas the optimal gap in previous studies never exceeded 1 day (Figure 1), presumably because the test delays in these studies never exceeded 1 week.

In an effort to formally describe this nonmonotonic effect of gap on final test score, we fit to these data a mathematical function that inherently produces the sharp ascent and gradual descent illustrated in Figures 3 and 4,

$$y = -a\,[ln(g + 1) - b]^2 + c.$$

This function expresses final test score (y) as a quadratic function of the natural logarithm of gap (g), which produces a positively-skewed downward-facing parabola with shape and position depending on the parameters a, b, and c. Although this function is not theoretically motivated, its parameters are meaningful. In particular, parameter c equals the optimal test score, and $e^b - 1$ equals the optimal gap. Fits of this function to the data in Experiments 1, 2 (facts), and 2 (objects) produced optimal test scores of 71%, 52%, and 21%, respectively, and optimal gaps of 3.7, 25.6, and 37.1 days, respectively. The function explained a moderate amount of variance (with $R^2 = .67, .90$, and $.75$, respectively). By contrast, the variance explained by a line ($R^2 = .004, .10$, and $.14$, respectively) was far less than that explained by numerous nonlinear functions with just two parameters.

Additional tentative conclusions can be reached. First, whereas an increase in gap from several minutes to the optimal gap produced a major gain in long-term retention, further increases in gap (from the optimal to the longest gap we tested) produced relatively small and nonsignificant (Experiment 1, $p = .463$; Experiment 2, $p = .448$) – but not trivial – decreases in both final recall and recognition. Thus, the penalty for a too-short gap is far greater than the penalty for a too-long gap. Second, by comparing the results of Experiment 1 (in which a 10-day test delay produced an optimal gap of 1 day) and Experiment 2 (in which a 6-month test delay produced an optimal gap of 1 month), one might conclude that optimal gap becomes larger as test delay gets larger. Because Experiments 1 and 2 used different materials and procedures, it is possible that the change in optimal gap could be due to those differences and not due to increased test delay. However, because previous studies have shown optimal gap invariance using a wide range of materials and procedures, we believe that the increase in optimal gap is truly related to increased test delay. The 6-month test delay experiment presented here suggests that a 1-day gap is far from optimal when the test delay is longer than 1 month. Just as short-test delay studies have demonstrated that optimal gap increases as test delay increases, these results tentatively indicate that the same holds true at long test delays.

Next, we consider our findings in relation to the literature. Figure 5 shows a log-log plot of optimal gap as a function of test delay, for every study in the Cepeda et al. (2006) meta-analysis containing an optimal gap, plus data from the present paper (total of $n = 48$ data points). Two features can be seen: First, optimal gap increases as a function of test delay. Second, the ratio of optimal gap to test delay appears to decrease as a function of test delay. At very short test delays, on the order of minutes, the ratio is close to 1.0; at multiday test delays, the ratio is closer to 0.1. These data are at odds with the notion that the optimal gap/test-delay ratio is independent of test delay, as some have speculated (Crowder, 1976; Murray, 1983). Instead, the present findings, in conjunction with the literature, are consistent with the possibility that the optimal gap increases with test delay, albeit as a declining proportion of test delay.[3]

Encoding variability theories, such as Estes' stimulus fluctuation model (Estes, 1955), hold that study context is stored along with an item, and itself changes with time. As gap increases, there is an increase in the expected difference between the encoding contexts occurring at each study episode. Similarity between encoding and retrieval contexts is assumed to result in a greater likelihood of recall (Glenberg, 1979), and spacing improves retention by increasing the chance that contexts during the first or second study episode will match the retrieval context, thereby increasing the probability of successful trace retrieval. Both a published encoding variability model (Raaijmakers, 2003) and our own preliminary modeling efforts (Mozer, Cepeda, Pashler, Wixted, & Rohrer, 2008) lend support to this theory. Alternatively, study-phase retrieval theories (Hintzman, Summers, & Block, 1975; Murray, 1983) propose that each time an item is studied, previous study instances are retrieved. To the extent that the retrieval process is both successful and increasingly difficult, increasingly large distributed practice effects should be observed. Study-phase retrieval theories predict – and our data show – an inverted-U-shaped function of gap on performance following a test delay.

[3] Because we only used a limited range of gaps in the present studies, and the true optimal gap in each of our studies might be slightly shorter or longer than the observed optimal gap, our current data neither support nor refute the existence of a further decreasing ratio, within the multiday test delay period.

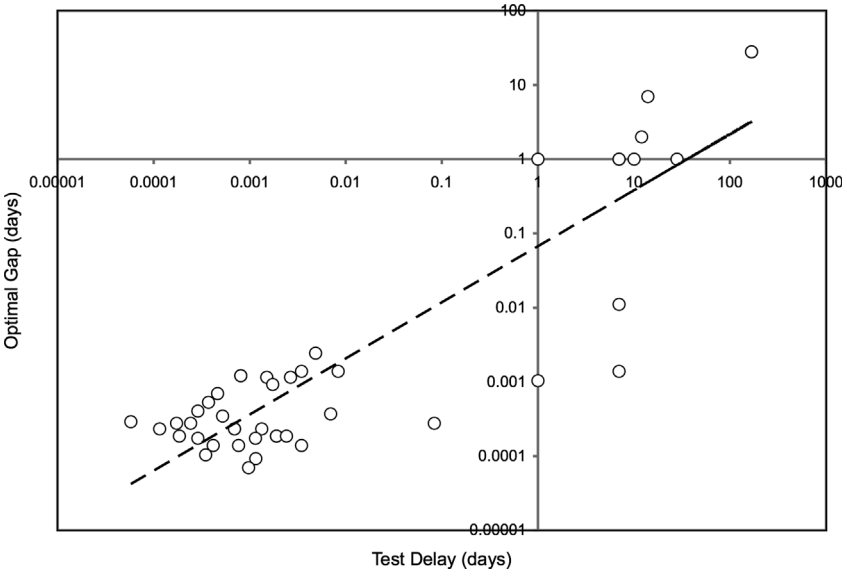

Figure 5. Log-log plot of optimal gap value by test delay, for all studies in the Cepeda et al. (2006) meta-analysis for which the optimal gap was flanked by shorter and longer gaps. The dashed line shows the best-fit power regression line for the observed data. Optimal gap increases as test delay increases, and the ratio of optimal gap to test delay decreases as test delay increases.

Practical Implications

To efficiently promote truly long-lasting memory, the data presented here suggest that very substantial temporal gaps between learning sessions should be introduced – gaps on the order of months, rather than days or weeks. If these findings generalize to a classroom setting – and we expect they will, at least with regard to learning "cut and dry" kinds of material – they suggest that a considerable redesign of conventional instructional practices may be in order. For example, regular use of cumulative tests would begin to introduce sufficiently long spacing gaps. Cramming courses and shortened summer sessions are especially problematic, as they explicitly reduce the gap between learning and relearning.

Failure to consider distributed practice research (Bahrick, 2005) is evident in instructional design and educational psychology texts, many of which fail even to mention the distributed practice effect (e.g., Bransford, Brown, & Cocking, 2000; Bruning, Schraw, Norby, & Ronning, 2004; Craig, 1996; Gardner, 1991; Morrison, Ross, & Kemp, 2001; Piskurich, Beckschi, & Hall, 2000). Those texts that mention the distributed practice effect often devote a paragraph or less to the topic (e.g., Glaser, 2000; Jensen, 1998; Ormrod, 2003; Rothwell & Kazanas, 1998; Schunk, 2000; Smith & Ragan, 1999) and offer widely divergent suggestions – many incorrect – about how long the lag between study sessions ought to be (cf. Gagné, Briggs, & Wager, 1992; Glaser, 2000; Jensen, 1998; Morrison et al., 2001; Ormrod, 2003; Rothwell & Kazanas, 1998; Schunk, 2000; Smith & Ragan, 1999).[4] The present studies begin to fill in the gaps that have maintained this unsatisfactory state of affairs and suggest the need for research that applies distributed practice principles within classrooms and embeds them within educational technologies.

Acknowledgments

Nicholas J. Cepeda, Department of Psychology, York University and Department of Psychology, University of California, San Diego; Harold Pashler, Noriko Coburn, and John T. Wixted, Department of Psychology, University of California, San Diego; Doug Rohrer, Department of

[4] Gagné et al. (1992) suggest reviewing the material after an interval of weeks or months; in fact, review, as compared to testing with feedback, is a poor way to restudy the information (Pashler et al., 2005). Additionally, Gagné et al. state that distributed practice improves concept learning but we have found no existing studies in the literature that support this claim, and our recent studies fail to support this claim (Pashler et al., 2007). Jensen (1998) suggests using 10 min, 2 day, and 1 week reviews of material; no empirical studies or theories would predict the spacing intervals cited to be ideal. Morrison et al. (2001) suggest writing facts over and over to learn them; this prescription for massed practice and overlearning is a highly inefficient use of time (Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005). Ormrod (2003) suggests distributing reviews over a period of months or years; the same caveats already mentioned, such as the relative ineffectiveness of review versus testing with feedback, apply here. Rothwell and Kazanas (1998) suggest reviewing material periodically; this is vague, and, again, review is not ideal. Schunk (2000) suggests spaced review sessions; the caveats already mentioned apply here. Smith and Ragan (1999) incorrectly claim that massed practice benefits association learning, when in fact most studies have shown that distributed practice improves memory for paired associates.

Psychology, University of South Florida; Michael C. Mozer, Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder.

This work was supported by the Institute of Education Sciences, US Department of Education, through Grants R305H020061 and R305H040108 to the University of California, San Diego, and by the US National Science Foundation (Grant BCS-0720375, H. Pashler, PI; and 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 US Department of Education. The authors are grateful to Kitu Jhawar for general assistance, to David M. Perlmutter for help in selecting the Swahili dialect, to Phil Starkovich and Grant Yoshida for programming the studies, to Ed Vul for assistance with data analysis, to Carol Armstrong for lending her voice to the fact and picture tasks, and to Phuong An, Abel Aramburo, Melissa Ares, Carol Armstrong, Shana Carr, Gilbert Franco, Rubie Garcia, Phuong Le, Rebeca Limon, Cathy Nylin, Steve Rodriguez, Krisvell Sanchez, Nick Thaler, Jonathan Wey, and Tony Ybarra for collecting the data. We thank John R. Anderson, Jeroen Raaijmakers, Stephan Lewandowsky, and Art Glenberg for helpful comments on a previous version of this manuscript.

A preliminary description of initial data from part of this study, which was written before the data collection was completed, was included in a review article summarizing our research program (Pashler et al., 2007).

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York: Academic Press.

Ausubel, D. P. (1966). Early versus delayed review in meaningful learning. *Psychology in the Schools, 3*, 195–198.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*, 296–308.

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.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research, 74*, 245–248.

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika, 26*, 255–280.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington, DC: National Academy Press.

Bruning, R. H., Schraw, G. J., Norby, N. M., & Ronning, R. R. (2004). *Cognitive psychology and instruction* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Burtt, H. E., & Dobell, E. M. (1925). The curve of forgetting for advertising material. *Journal of Applied Psychology, 9*, 5–21.

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.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology, 38*, 967–978.

Craig, R. L. (1996). *The ASTD training and development handbook: A guide to human resource development* (4th ed.). New York: McGraw-Hill.

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.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1*, 309–330.

Dempster, F. N., & Perkins, P. G. (1993). Revitalizing classroom assessment: Using tests to promote learning. *Journal of Instructional Psychology, 20*, 197–203.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect. *Journal of Applied Psychology, 84*, 795–805.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications, Inc. (Original work published in 1885).

Edwards, A. S. (1917). The distribution of time in learning small amounts of material. In *Studies in psychology, contributed by colleagues and former students of Edward Bradford Titchener* (pp. 209–213). Worcester, MA: Louis N. Wilson.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62*, 369–377.

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology, 59*, 290–296.

Gagné, R. M., Briggs, L. J., & Wager, W. W. (1992). *Principles of instructional design* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

Gardner, H. (1991). *The unschooled mind: How children think and how schools should teach*. New York: Basic Books.

Glaser, R. (2000). *Advances in instructional psychology* (Vol. 5, Educational design and cognitive science). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Gordon, K. (1925). Class results with spaced and unspaced memorizing. *Journal of Experimental Psychology, 7*, 337–343.

Harzem, P., Lee, I., & Miles, T. R. (1976). The effects of pictures on learning to read. *British Journal of Educational Psychology, 46*, 318–322.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 31–40.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30*, 138–149.

Jensen, E. (1998). *Teaching with the brain in mind*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jost, A. (1897). Die Assoziationsfestigkeit in ihrer Abhängigkeit von der Verteilung der Wiederholungen [The strength of associations in their dependence on the distribution of repetitions]. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane, 14*, 436–472.

Keppel, G. (1964). Facilitation in short- and long-term retention of paired associates following distributed practice in learning. *Journal of Verbal Learning and Verbal Behavior, 3*, 91–111.

Morrison, G. R., Ross, S. M., & Kemp, J. E. (2001). *Designing effective instruction* (3rd ed.). New York: John Wiley & Sons.

Moss, V. D. (1995). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student (Doctoral dissertation, Utah State University, 1995). *Dissertation Abstracts International 56*, 5204.

Mozer, M. C., Cepeda, N. J., Pashler, H., Wixted, J. T., & Rohrer, D. (2008). *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.

Ormrod, J. E. (2003). *Educational psychology: Developing learners* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate retention of words? *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 3–8.

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

Peigneux, P., Laureys, S., Delbeuck, X., & Maquet, P. (2001). Sleeping brain, learning brain. The role of sleep for memory systems. *NeuroReport, 12*, A111–A124.

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

Piskurich, G. M., Beckschi, P., & Hall, B. (Eds.). (2000). *The ASTD handbook of training design and delivery: A comprehensive guide to creating and delivering training programs – instructor-led, computer-based, or self-directed*. New York: McGraw-Hill.

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

Robinson, E. S. (1921). The relative efficiencies of distributed and concentrated study in memorizing. *Journal of Experimental Psychology, 4*, 327–343.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19*, 361–374.

Rose, R. J. (1992). Degree of learning, interpolated tests, and rate of forgetting. *Memory & Cognition, 20*, 621–632.

Rothwell, W. J., & Kazanas, H. C. (1998). *Mastering the instructional design process: A systematic approach* (2nd ed.). San Francisco: Jossey-Bass.

Schunk, D. H. (2000). *Learning theories: An educational perspective* (3rd ed.). Upper Saddle River, NJ: Merrill.

Shuell, T. J. (1981). Distribution of practice and retroactive inhibition in free-recall learning. *The Psychological Record, 31*, 589–598.

Simon, J. L. (1979). What do Zielske's real data really show about pulsing? *Journal of Marketing Research, 16*, 415–420.

Smith, P. L., & Ragan, T. J. (1999). *Instructional design* (2nd ed.). Upper Saddle River, NJ: Merrill.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641–657.

Strong, E. C. (1973). The effects of repetition in advertising: A field experiment (Doctoral dissertation, Stanford University, 1972). *Dissertation Abstracts International, 33*, 4615.

Strong, E. K. Jr. (1916). The factors affecting a permanent impression developed through repetition. *Journal of Experimental Psychology, 7*, 319–338.

Watts, D., & Chatfield, D. (1976). Response availability as a function of massed and distributed practice and list differentiation. *The Psychological Record, 26*, 487–493.

Welborn, E. L. (1933). A study of logical learning in college classes. 20th Annual Conference on Educational Measurement. *I. U. School of Education Bulletin, 10*, 12–20.

Received December 17, 2007
Revision received May 8, 2008
Accepted May 9, 2008

Nicholas J. Cepeda

York University
Department of Psychology
4700 Keele Street
Toronto
ON M3J 1P3
Canada
E-mail ncepeda@yorku.ca

