APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **23**: 1297–1317 (2009)
Published online 19 January 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1537

Optimising Learning Using Flashcards: Spacing Is More Effective Than Cramming

NATE KORNELL*

Department of Psychology, University of California, Los Angeles, USA

SUMMARY

The spacing effect—that is, the benefit of spacing learning events apart rather than massing them together—has been demonstrated in hundreds of experiments, but is not well known to educators or learners. I investigated the spacing effect in the realistic context of flashcard use. Learners often divide flashcards into relatively small stacks, but compared to a large stack, small stacks decrease the spacing between study trials. In three experiments, participants used a web-based study programme to learn GRE-type word pairs. Studying one large stack of flashcards (i.e. spacing) was more effective than studying four smaller stacks of flashcards separately (i.e. massing). Spacing was also more effective than cramming—that is, massing study on the last day before the test. Across experiments, spacing was more effective than massing for 90% of the participants, yet after the first study session, 72% of the participants believed that massing had been more effective than spacing. Copyright © 2009 John Wiley & Sons, Ltd.

The spacing effect—that is, the fact that spacing learning events apart results in more long-term learning than massing them together—is a robust phenomenon that has been demonstrated in hundreds of experiments (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1996; Hintzman, 1974; Glenberg, 1979) dating back to Ebbinghaus (1885/1964). Learners would profit from taking advantage of the spacing effect, both in classrooms and during unsupervised learning (e.g. Bahrick, Bahrick, Bahrick, & Bahrick, 1993)—and doing so seems feasible from a practical perspective because spacing does not take more time than massing, it simply involves a different distribution of time (Rohrer & Pashler, 2007). Yet the spacing effect has been called 'a case study in the failure to apply the results of psychological research' (Dempster, 1988, p. 627). One reason for this failure is that spacing has seldom been investigated using procedures that are directly applicable in educational settings (although there are exceptions, e.g. Rohrer & Taylor, 2006, 2007; Smith & Rothkopf, 1984). For example, in spacing experiments, a spaced condition is often compared to a pure massing condition, in which the same item (e.g. a word pair) is presented twice in a row with no intervening items. Pure massing is ineffective, but it is also often unrealistic (Seabrook, Brown, & Solity, 2005). The goals of the present experiments were twofold: First, to investigate the spacing effect in a realistic study situation, and second, to examine students' attitudes towards spacing as a study strategy. The research was also intended to provide learners with practical information about how to study.

*Correspondence to: Nate Kornell, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095, USA. E-mail: nkornell@ucla.edu

The present experiments were modelled on flashcards, which are among the most common tools that learners use to study facts (Kornell & Bjork, 2008b). When studying with flashcards, learners can make a variety of decisions. One such decision is: To optimise learning efficiency, how many flashcards should one include in a flashcard stack at one time? This decision influences the spacing between study trials. The larger the stack, the larger the *within-session spacing*—that is, the larger the spacing between repetitions of a given card. For example, in a stack of 20 cards, repetitions of a given card are separated by 19 other cards. In a stack of five cards, by contrast, only four cards intervene between repetitions of a given card. Another decision that learners face is how much (if any) spacing to allow between study sessions—that is, *between-session spacing*. For example, studying a stack of flashcards four times in a row on a single day results in less spacing than studying the same stack for the same total amount of time, but on four different days.

Learners' attitudes towards spacing

Learners often base decisions about what, when and how to study on metacognitive judgments they make about their own memories (Kornell & Bjork, 2007; Kornell & Metcalfe, 2006; Nelson, Dunlosky, Graf, & Narens, 1994). When learners make decisions about how many flashcards to study at a given time; however, they may not consider the impact of their decisions on spacing. Learners frequently neglect the effects of spacing when making study decisions (Kornell & Bjork, 2007). When they do consider spacing, they often exhibit the illusion that massed study is more effective than spaced study, even when the reverse is true (Dunlosky & Nelson, 1994; Kornell & Bjork, 2008a; Simon & Bjork, 2001; Zechmeister & Shaughnessy, 1980). (There is also evidence that learners chose to space or mass differentially depending on whether the to-be-learned materials are easy or difficult; Benjamin & Bird, 2006; Son, 2004.) One explanation for the illusion that massing is effective is that massing makes studying seem easier and faster than does spacing (Baddeley & Longman, 1978). In studying flashcards, for example, learners tend to test themselves as they study, by looking at the question posed on the front of a card and trying to recall the answer before turning the card over to reveal the answer. These recall attempts have a large influence on the judgments people make about their own memories (e.g. Dunlosky & Nelson, 1992; Finn & Metcalfe, 2008; Koriat, 1997; Spellman & Bjork, 1992). Moreover, such recall attempts are easier if one chooses to mass learning (i.e. use small stacks of flashcards or short intervals between study sessions) than if one chooses to space learning. Thus spacing can reduce performance levels during learning and simultaneously enhance long-term learning. When people make the mistake of assuming that short-term performance equals long-term learning—which they often do (Bjork, 1994, 1999)—they may convince themselves that as a study strategy, massing is more effective than spacing.

The advice students receive about how to study effectively—if they receive any, which is uncommon (Kornell & Bjork, 2007; Son & Kornell, 2008)—seems to be misguided with respect to the benefits of spacing. The initial impetus for the present experiments came from a leading Graduate Record Examination (GRE)[1] study guide that advises students to use relatively small stacks of flashcards, as well as from the recommendation of researchers who advocate using small stacks of flashcards (e.g. Salisbury & Klein, 1988). Moreover, learners often choose to use small stacks; for example, Salisbury and Klein (1988) found that when they gave participants a large stack of flashcards, 83% spontaneously divided the

[1] The Graduate Record Examination is a graduate school entrance examination.

DOI: 10.1002/acp

cards into smaller stacks. Relatively small stacks of flashcards may be advantageous for reasons of motivation and convenience. In the present experiments, I tested the hypothesis that using small stacks of flashcards would have a negative impact on learning and memory.

The ultimate example of massing is—and perhaps the most common study technique of all—is cramming. Cramming involves studying something intensely, often for the first time, in the days or hours before a test. Procrastination followed by cramming is popular among students (e.g. Brinthaupt & Shin, 2001) and equally unpopular among educators (with the possible exception of teachers preparing a class for an upcoming high-stakes standardised test). Cramming is the opposite of spacing, but cramming has its own advantages (Vacha & McBride, 1993). The main advantage of cramming in the present context is that there is little time for forgetting between the time of study and the time of the test. Thus cramming may provide a boost to students' test scores and grades by creating short-lived memories without creating the type of lasting memories that are the ultimate goal of education. Experiments on the spacing effect do not necessarily address the issue of cramming because the massed condition does not always occur just before the test. In the present experiments, I examined cramming in two ways: First, by comparing cramming (i.e. massed study immediately prior to a test) to an equivalent amount of spaced study, and second, by examining the effect of providing a review session, in which students were given a chance to review everything that they had learned, before the test.

The present experiments

In the present experiments, I compared two strategies students use when they study with flashcards. The spaced condition corresponded to the strategy of using a big stack of 20 flashcards; the massed condition corresponded to the strategy of splitting flashcards into four smaller stacks of five cards each. In Experiment 1, to investigate the effects of within-session spacing, participants studied flashcards in one large stack (i.e. spaced) or in four smaller stacks (i.e. massed). Between session spacing was held constant. To increase the realism of Experiment 2, participants studied the flashcards across four days; in the spaced condition, they studied the same large stack each day, whereas in the massed condition they studied a different small stack each day. The test occurred on the fifth day. Items studied in the massed condition in the fourth study session were used to evaluate the relative benefits of cramming. To increase realism further in Experiment 3, after participants studied for four days, there was a day five review session, covering all of the to-be-learned materials, followed by a final test on day six. Because of the increasing popularity of online vocabulary learning, the experiments were conducted using a web-based study programme, instead of in the laboratory. In all three experiments, participants studied 40 GRE-type flashcards, 20 in the spaced condition and 20 in the massed condition.

EXPERIMENT 1

Experiment 1 investigated participants' ability to learn GRE-type vocabulary using flashcards. Between-session spacing was held constant, and only within-session spacing (i.e. the number of cards in a stack) was manipulated. The massed condition did not involve presenting the same item on consecutive trials, in contrast to most previous research, because consecutive presentations would not have reflected realistic flashcard use. There were two conditions: In the massed condition, participants studied four small stacks of

DOI: 10.1002/acp

flashcards separately, one stack at a time. In the spaced condition, participants studied one large stack of flashcards. Items from both conditions were presented for study during the same session in this experiment and the experiments that follow. Each flashcard was studied four times regardless of condition. The study phase took place online during a single session. The test phase, which was also online, occurred an average of 24 hours later.

Method

Participants

The participants were 20 University of California, Los Angeles (UCLA) students who participated in exchange for course credit.

Design

There were two within-participant conditions: In the spaced condition, word pairs were studied in a single large 'stack' of 20 cards; in the massed condition, the word pairs were split into 4 smaller stacks of 5 cards each.

Materials

The materials were 40 synonyms (e.g. '*effulgent*: *brilliant*') that were selected because they were typical of the types of words that appear on standardised tests such as the GRE (see Appendix). The goal, in creating the stimuli, was to select synonyms such that most participants would know the meaning of the second, but not the first, word in each pair. A pilot study was conducted to determine whether this goal was accomplished. Participants' knowledge of each word was tested by presenting the word accompanied by five definitions, one of which was correct, in a multiple-choice format. The same set of five response options was used for both synonyms in a given pair; for example, the correct response for both effulgent and brilliant was *shining brightly; radiant*. Because a given participant could not be tested on both synonyms in a pair, two 40-item tests were created, each containing 20 cues (i.e. first words) and 20 targets (i.e. second words). The same set of 40 five-item multiple choice responses were used for both tests. Each version of the test was given to 12 participants. The results showed that, as expected, participants selected the correct definition for cues (M = .39, SD = .20) less frequently than they selected the correct definition for targets (M = .95, SD = .06). Statistical analyses of test accuracy, which compared each set of 20 cues to its corresponding set of 20 targets on a between-participant basis, demonstrated that the difference was significant for both item sets, $t(22) = 6.63$, $p < .0001$, and $t(22) = -14.27$, $p < .0001$.

Setting

The experiments presented in this article all took place online. Allowing participants to participate from home, or wherever they were, instead of in the lab, increased the realism and generalisability of the experiments.

Procedure

At the outset of the experiment, participants were told that they would be studying digital flashcards. They were told they could spend as long as they wanted studying each word, and that after they finished studying there would be a cued recall test, on which the first word in each pair would be presented and they would be asked to type in the second word. The instructions described the procedure in detail, but to avoid disrupting participants' natural study patterns, there were no specific directions about how to study (e.g. 'test yourself while looking at the "front" of each flashcard'). There were two sessions; session 1 consisted of a learning phase, and session 2 consisted of a cued-recall test.

During session 1 (i.e. the learning phase), there were two conditions. In the spaced condition, a single set of 20 word pairs (i.e. a large stack) was presented for study. The entire stack was presented, always in the same order, four consecutive times. In the massed condition, four different sets of five word pairs were presented (i.e. the small stacks). Each stack was presented four consecutive times, always in the same order, before the next stack was presented. Thus in both conditions, every word pair was presented four times. A given participant was randomly assigned to study either the massed words first followed by the spaced words, or vice versa.

On each trial, a word pair was presented on a computer using a web-based study programme. First, the cue word was presented, followed by a blank (e.g. 'effulgent: ________'). The word remained visible until the participant pressed a button labelled 'next' (i.e. the participant controlled the timing of the presentations). Then the cue disappeared and the target appeared preceded by a blank (e.g. '________: brilliant'). Again, the participant controlled the timing of the presentation. In an effort to stay true to real-life flashcards, which do not require overt responses, participants were not explicitly tested, or required to provide answers of any kind, during the study phase.

Word pairs were randomly assigned to either the spaced or massed condition on a participant-by-participant basis. The order of the words was also assigned randomly, although once assigned, it remained fixed for a given participant.

The final test took place during session 2. During the final test, participants were tested on all 40 definitions, one by one, in random order. The first word in each pair was presented (e.g. effulgent) and participants were asked to type in the synonym (e.g. brilliant). Participants were asked to take the final test approximately 24 hours after they finished studying, although when they took it was ultimately their decision. The median delay between study and test was 24 hours, and the range was 17–41 hours.

At the end of each session, participants were asked to make separate estimates, for the massed and spaced conditions, of how much they had learned. At the end of session 1, participants were asked the following questions: 'All of the flashcards that you studied were presented 4 times. You may have noticed that there were two conditions. Massed cards were repeated once every 5 cards. Spaced cards were repeated once every 20 cards. What percentage of the massed words do you think you will be able to remember on the test tomorrow? What percentage of the spaced words do you think you will be able to remember on the test tomorrow?' At the end of session 2, participants were asked a similar question, reworded in the past tense.

Results and discussion

In all of the experiments presented here, test responses were scored using a computer algorithm that counted correctly spelled and misspelled correct answers as correct. The percentage of items recalled was significantly higher in the large stack (spaced) condition (M = 49%, SD = 27) than in the small stack (massed) condition (M = 36%, SD = 26), $t(19) = 2.26$, $p < .05$, $d = .48$.

Study time

The amount of time spent studying was under the participants' control. To measure study time, I computed the total time spent studying the cue and the target across all trials on a given item. Outliers, which were defined as study time scores more than two standard deviations away from the mean, were excluded from the analyses. Mean study time scores were then computed for each participant, separately for the massed and spaced conditions. The mean number of seconds participants spent studying in the spaced condition (M = 22.60, SD = 10.01) and the massed condition (M = 22.19, SD = 10.74) did not differ significantly, $t(19) = .22$, $p = .83$, $d = .04$. A second analysis, in which outliers were not excluded, also produced a non-significant result.

To investigate further the effectiveness of spacing, a study efficiency metric was computed by dividing the number of correct responses by the number of minutes spent studying in each condition (see Pyc & Rawson, 2007). Study efficiency (i.e. the number of items learned per minute) in the spaced condition (M = 1.38, SD = .74) and the massed condition (M = 1.15, SD = .94) did not differ significantly, $t(19) = 1.37$, $p = .19$, $d = .28$.

Judgments of learning

The estimates participants made, at the end of session 1, of the number of items they would recall on the session 2 test were significantly lower in the spaced condition (M = 43%, SD = 31) than the massed condition (M = 50%, SD = 29), $t(19) = 2.35$, $p < .05$, $d = .23$—contrary to actual recall performance. At the end of session 2, after the test, participants' estimates of the number of items recalled did not differ between the spaced condition (M = 26%, SD = 27) and the massed condition (M = 28%, SD = 25), $t(19) = .57$, $p = .57$. However, the most important estimates may be the ones made at the end of session 1, because, for students who do not regularly experiment with how they study, initial impressions seem likely to control subsequent study decisions.

In summary, participants in Experiment 1 learned more in the large-stack (spaced) condition than they did in the small-stacks (massed) condition. Nonetheless, after experiencing both the massed and spaced conditions, participants believed that massed study was more effective than spaced study. These findings suggest that even in the absence of between-session spacing, within-session spacing enhances learning, and that while using small stacks of flashcards may be popular, it is detrimental to learning; they also suggest that using small stacks of flashcards creates an illusion of effective learning

EXPERIMENT 2

Experiment 2 was designed to compare the same two conditions as Experiment 1, but in a more realistic situation. In reality, learners tend to study the same material on different days (except for the worst procrastinators, who study only at the last minute), and often in a variety of temporal and physical settings. Thus in Experiment 2, each of the four study sessions took place on a different day, and the test took place on the fifth day.

Method

The materials, design and setting of Experiment 2 were the same as they were in Experiment 1.

Participants

The participants were 25 UCLA students who participated in exchange for course credit.

DOI: 10.1002/acp

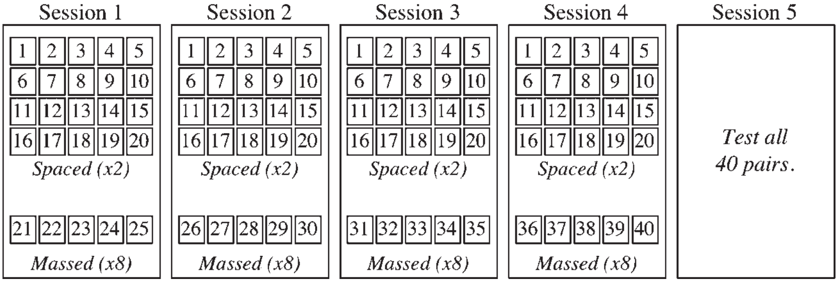

Figure 1. Experiment 2 procedure. During every session, 20 spaced items were presented two times each, and five massed items were presented eight times each. The same set of 20 spaced items (denoted as items 1–20 here) were studied every session; a different set of five massed items (denoted as items 21–40 here) were studied every session. The cued-recall test occurred during the fifth session. In Experiment 3, the only procedural change was that during session 5, all items from both conditions were presented two times each for review, and the test occurred during session 6

Procedure

The procedure of Experiment 2 was similar to the procedure of Experiment 1. The principal change was that each session took place on a different day. Four days of study were followed by a test on the fifth day (see Figure 1). Thus between-session spacing (i.e. the amount of time between study sessions) and within-session spacing (i.e. number of cards intervening between study trials on any particular card) were both manipulated, in a way that was intended to be a realistic simulation of actual studying. The spaced items comprised a single stack of 20 flashcards, which participants studied twice during each of the four study sessions. The massed items were split into four stacks, and participants studied each stack eight times on a single day. The number of study opportunities was increased to eight in Experiment 2, from four in Experiment 1, to insure that items were studied multiple times each day in both the massed and spaced conditions.

Participants were asked to participate on five consecutive days if possible, and they were asked not to skip more than one day, although they ultimately decided when to participate. The median time between any two consecutive sessions was 24 hours, and the range was 10–63 hours.

At the end of each session, participants were asked to predict how well they would do on the final test in each condition. To avoid confusion during the predictions, I emphasised the fact that there would be four study sessions and that the massed items would each occur in a single session whereas the spaced items would be repeated in every session. The exact questions were as follows:

> You have just studied two separate word lists. As you have probably noticed, one set of words was repeated multiple times today, whereas you saw another set of words only once. The set of words that was repeated is called the ‘massed’ set. You will see a set of different repeated words like this on every day of practice. Together, all of these repeated words represent the ‘massed list.’ On the other hand, the words that you only saw once today are known as the ‘spaced’ set. You will see this exact same list exactly one time on every day of practice. This is called the ‘spaced list.’ In total, you will see 40 words in the massed list and another 40 in the spaced list. On the final day you will be tested to see how many words from each list you can remember. What percentage of

DOI: 10.1002/acp

> the words from the massed list do you think you will be able to remember? What percentage of the words from the spaced list do you think you will be able to remember?''

Due to an experimenter error, the instructions described spaced items—which were studied two times per session—as having been studied once per session. However, it is unlikely that this error caused significant confusion on the participants' parts, because, consistent with the instructions, the massed condition included more presentations of a given item than did the spaced condition. (Note that because there were 20 word pairs per condition, the instructions referred to each condition as including 40 words.)

Results and discussion

Like Experiment 1, memory performance was better in the spaced condition (M = 54%, SD = 35) than the massed condition (M = 21%, SD = 19), $t(24) = 6.03, p < .0001, d = 1.18$ (see Figure 2). The difference in performance between the spaced and massed conditions was larger in Experiment 2 than Experiment 1, perhaps because between-session spacing was manipulated.

Experiment 2 allowed an examination of the effects of cramming. Items studied in the massed condition during the final study block were categorised as cramming items. The cramming items were remembered at a high rate relative to other massed items, as shown by a main effect of study session in the massed condition, $F(3, 72) = 5.12$, $p < .01$, $\eta_p^2 = .18$. Nonetheless, a planned comparison revealed that spacing (M = 54%, SD = 35) was more effective than cramming (i.e. massed study during session 4, M = 34%, SD = 36), $t(24) = 2.77$, $p < .05$, $d = .55$. This effect occurred despite the fact that there were only five to-be-learned items in the cramming condition, whereas there were 20 to-be-learned items in the spaced condition.

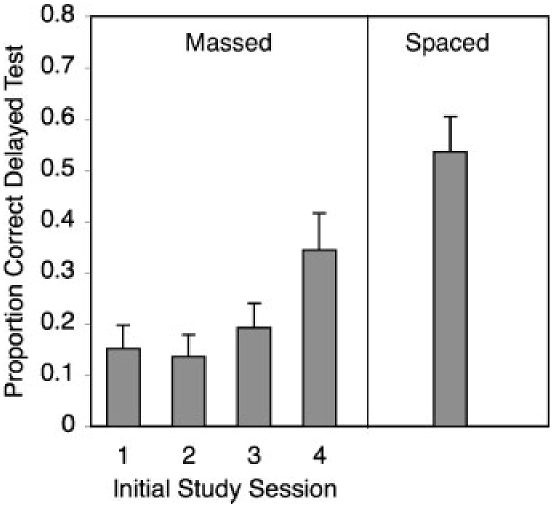

Figure 2. Proportion correct on the delayed test as a function of spacing condition and the session during which the pair was first studied in Experiment 2. All of the spaced items were first studied in session 1. The massed items studied in session 4 represent cramming. The test took place during session 5. Error bars represent 1 SEM

DOI: 10.1002/acp

Study time

Again, study time scores more than two standard deviations away from the mean were excluded from the analyses. The mean number of seconds participants spent studying in the spaced condition (M = 40.30, SD = 21.95) and the massed condition (M = 42.94, SD = 26.36) did not differ significantly, *t*(24) = .76, *p* = .45, *d* = .11. There was also no significant difference in a second analysis, in which outliers were not excluded.

The study time data from Experiment 2 are displayed in Figure 3. There was a sharp decrease in the time participants spent studying across the eight study trials, from roughly 8 seconds on trial 1 (i.e. the first time an item was studied) to roughly 3.5 seconds on trial 8 (i.e. the last time an item was studied). Most of the study trials were completed fairly rapidly, considering that study time on a given trial was defined as the sum of the time spent studying the cue and the target. (A similar pattern of study time occurred in Experiment 1.)

Unlike Experiment 1, an efficiency analysis showed that participants learned significantly more items per minute of study in the spaced condition (M = .81, SD = .49) than the massed condition (M = .34, SD = .35), *t*(24) = 3.86, *p* < .001, *d* = 1.11.

Judgments of learning

At the end of session 1, participants estimated that they had learned more in the massed condition (M = 60%, SD = 27) than the spaced condition (M = 41%, SD = 30), *t*(24) = 3.55, *p* < .01, *d* = .67. There were no significant differences between massed and spaced estimates in sessions 2, 3, or 4 (all *t*'s < 1). Again, the impressions learners form during their first study session are probably of primary importance, because those impressions are likely to serve as the basis for subsequent study decisions.

To summarise, Experiment 2 demonstrated, in a fairly realistic study situation, that studying one relatively large set of flashcards over a period of days (i.e. spacing) was superior to concentrating on a separate set of flashcards each day (i.e. massing). Spacing was also superior to cramming (i.e. studying intensively, eight times, during the final study

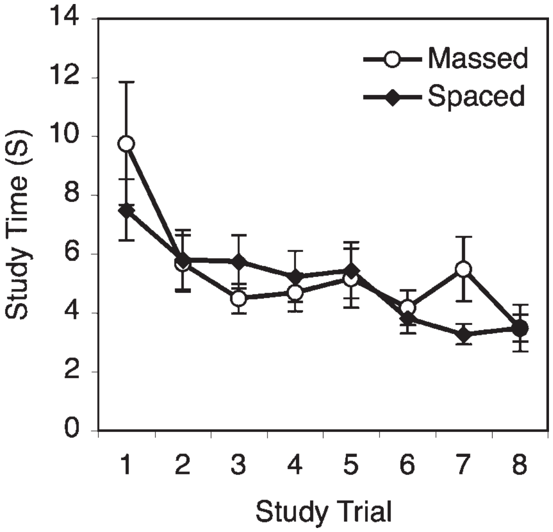

Figure 3. Mean study time per pair (including time on the cue and target) in Experiment 2, as a function of spacing condition and study trial. Error bars represent 1 SEM

DOI: 10.1002/acp

session). Despite the benefits of spacing, participants rated massing as more effective than spacing after experiencing both conditions.

EXPERIMENT 3

Experiment 3 was designed to address a shortcoming that Experiments 1 and 2 had in common with most previous experiments on the spacing effect: There was no final review session. In reality, before taking a test, virtually all students spend time reviewing the information that they will be tested on. The first two experiments lacked such a review. In Experiment 2, for example, the items participants studied in session 1 were not exposed again until the test during session 5. In Experiment 3, a review session was added to the procedure after the fourth study session but before the test. During the review session, all items from both conditions were presented for study twice. The review session test took place during session 5, and the test was moved to session 6.

The purpose of the review session was to make the procedure more realistic, and thus more generalisable to actual study situations. The review session had two additional consequences, however, both of which favoured the massed condition. First, the review session insured that the average delay from the final study trial to the test was the same in the massed and spaced conditions—all of the items were studied for the last time during session 5. Second, and more important, because of the review session, all of the ‘massed’ items were actually spaced between sessions. That is, they were presented for study in two different sessions—during their original study session and during the review session. These two features of Experiment 3 were both expected to benefit massed items more than spaced items.

Method

The materials, design and setting of Experiment 3 were the same as they were in Experiment 1 and 2.

Participants

The participants were 25 UCLA undergraduates who participated online. The median time between any two consecutive sessions was 25 hours and 30 minutes, and the range was 8–77 hours.

Procedure

The procedure was similar to the procedure of Experiment 2 (see Figure 1). The main procedural change was that a fifth session was inserted before the test, and the test took place during session 6. During the fifth session, all of the pairs were presented, twice each, for review. The large (i.e. spaced) stack was presented twice, just as it had been in the previous sessions. Each massed pair was also presented twice. The massed pairs were presented in the same order as they had been presented in during the previous sessions (i.e. the words from session 1, then session 2, then session 3, then session 4), but all 20 massed items were presented consecutively. After being presented a first time, the entire set was presented a second time, like in the spaced condition. Study trials in session 5 were identical to study trials in the previous sessions (i.e. the cue was presented followed by the target, with timing under the participant's control). The test took place during the sixth session.

At the end of every session participants were asked to estimate how well they would do on the final test, separately for massed and spaced items. In Experiment 2 there was an error in the instructions; the error was fixed in Experiment 3. Again participants were reminded of the procedure before making their predictions, and in particular it was emphasised that, aside from the review session, the massed items would be presented in only one session whereas the spaced items would be repeated in every session. The exact instructions were as follows:

> You have just studied two separate word lists. As you have probably noticed, one set of words was repeated 8 times today (i.e., the Massed set), whereas you saw another set of words only twice each (i.e., the Spaced set). There will be 4 different sets of Massed words. You will study a different set each day. There will only be one spaced set. You will study it every day. In total, you will see 20 words in the massed set and another 20 in the spaced set. The massed items will be shown 8 times on a single day; the spaced, 2 times per day for 4 days. On the 5th day you will review all of the words from both lists. Then on the final day you will be tested to see how many words from each list you can remember. What percentage of the words from the massed lists do you think you will be able to remember? What percentage of the words from the spaced lists do you think you will be able to remember?

Because there were 20 word pairs per condition, the instructions referred to each condition as including 20 words.

Results

Final test accuracy was significantly higher in the large-stack (i.e. spaced) condition (M = 65%, SD = 28) than the small-stack (i.e. massed) condition (M = 34%, SD = 28) $t(24) = 6.32$, $p < .0001$, $d = 1.11$ (Figure 4). Thus even in the presence of a review session, spacing using large flashcard stacks was more effective than massing using small flashcard stacks.

Test accuracy was significantly greater in the large-stack (i.e. spaced) condition (M = 65%, SD = 28) than it was for massed items initially studied during session 4 (M = 37%, SD = 34), as a planned comparison showed ($t(24) = 4.93$, $p < .0001$, $d = .90$). In Experiment 2, the massed items studied during session 4 were categorised as cramming items; in the present experiment, however, session 4 was not analogous to cramming because of the review session during session 5. Unlike Experiment 2, in the massed condition, the session in which an item was first studied did not have a significant effect on final test performance ($F < 1$).

I also compared the large-stack (i.e. spaced) items (M = 65%, SD = 28) to the massed items initially studied during session 1 (M = 38%, SD = 38), because the total between-session spacing in the two conditions was the same (i.e. study began in session 1 and ended in session 5). A planned comparison showed that the advantage of spaced condition was significant, $t(24) = 3.46$, $p < .01$, $d = .82$.

Study time

Study time scores more than two standard deviations away from the mean were excluded from the analyses, as in the previous experiments. The mean number of seconds

DOI: 10.1002/acp

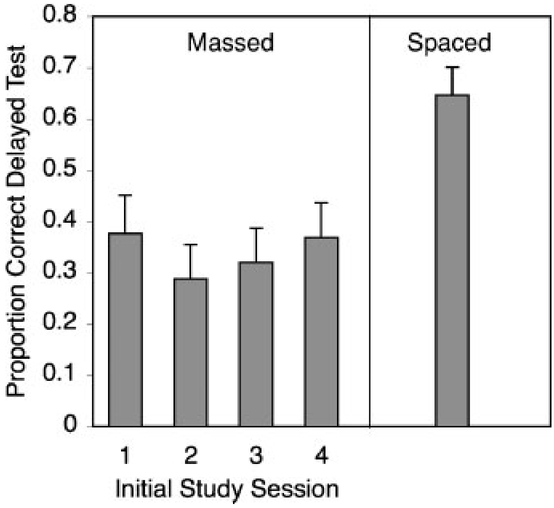

Figure 4. Proportion correct on the delayed test as a function of spacing condition and the session during which the pair was first studied in Experiment 3. All of the spaced items were first studied in session 1. All items were reviewed during session 5 and tested during session 6. Error bars represent 1 SEM

participants spent studying in the spaced condition (M = 41.47, SD = 17.67) and the massed condition (M = 43.78, SD = 19.94) did not differ significantly, $t(24) = .59$, $p = .56$, $d = .12$. Again, the same pattern was obtained when outliers were not excluded from the analyses.

The study time data from Experiment 3 are displayed in Figure 5. The pattern of results paralleled the results of Experiment 2: There was a sharp decrease in the time participants spent studying, from roughly 8 seconds on trial 1 to roughly 3 seconds on trial 10. Most of the study trials were completed fairly rapidly, considering that study time on a given trial was defined as the sum of the time spent studying the cue and the target.

An efficiency analysis showed that participants learned significantly more per minute of study in the spaced condition (M = .98, SD = .50) than the massed condition (M = .50, SD = .49), $t(24) = 4.93$, $p < .0001$, $d = .97$.

Judgments of learning

In contrast to actual test accuracy, the participants predicted, at the end of session 1, that they would do better on the final test on items that they had studied in the massed condition (66%) than items that they had studied in the spaced condition (51%). A planned comparison showed the difference to be significant, $t(23) = 2.25$, $p < .05$, $d = .58$ (one participant, who did not make performance estimates, was excluded from this analysis). The students seemed to learn from experience, however, and their ratings of the massed and spaced conditions did not differ significantly during sessions 2, 3 or 4 (all t's < 1). During session 5, participants rated massing (47%) as less effective than spacing (59%)—perhaps because returning to massed items from previous sessions made the participants recognise their inability to recall the massed items—although the difference only approached statistical significance, $t(23) = -1.67$, $p = .11$, $d = .47$.

DOI: 10.1002/acp

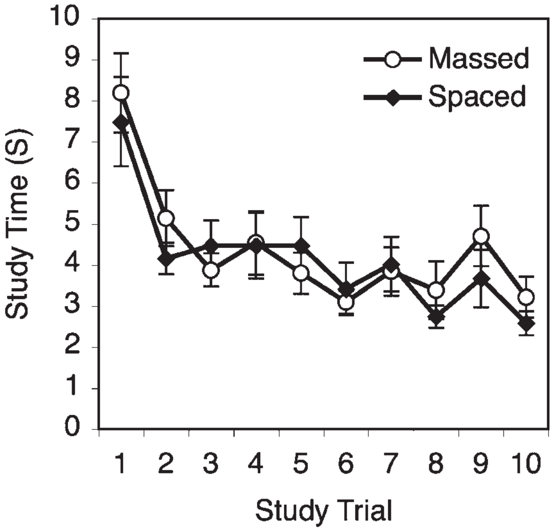

Figure 5. Mean study time per pair (including time on the cue and target) in Experiment 3, as a function of spacing condition and study trial. In the massed condition, eight of the trials took place in the same session, followed by two trials in the review session; in the spaced condition, two trials took place in each of the five sessions. Error bars represent 1 SEM

Combined analyses of Experiments 2 and 3

The effects of the review session can be assessed by comparing the results of Experiment 2 to the results of Experiment 3. It should be noted that participants in both experiments came from the same participant pool, but Experiment 2 was completed before Experiment 3 began, and thus participants were not assigned to experiments randomly.

In the combined analysis, spaced study was more effective than massed study ($F(1, 48) = 75.77$, $p < .0001$, $\eta_p^2 = .61$). Recall accuracy was higher in Experiment 3 (34% and 65% in the massed and spaced conditions, respectively) than in Experiment 2 (21% and 54%, respectively)—perhaps because of the review session in Experiment 3—although the difference was only marginally significant ($F(1, 48) = 2.98$, $p = .09$, $\eta_p^2 = .06$).

I had predicted that the review session would provide more benefit to the massed condition than the spaced condition, for two reasons: First, the 'massed' items were studied in multiple, spaced sessions in Experiment 3 (i.e. they were studied in their original study session and again, on a different day, during the review session), but in Experiment 2 they were only studied in one session. Second, in Experiment 3, unlike in Experiment 2, the lag from an item's final presentation to the test was equated. Contrary to these predictions, the review session did not diminish the size of the spacing effect. The advantage of spacing over massing was 33 percentage points in Experiment 2 and 31 percentage points in Experiment 3, and the experiment X spacing interaction did not approach significance ($F(1, 48) = .09$, $p = .77$).

Examining the massed items from Experiments 2 and 3 (Figures 2 and 4), it is apparent that the earlier an item was studied in Experiment 3, the more it benefited from review. Items studied in session 1 benefited the most from review, whereas items studied in session 4 were recalled equally in the two experiments. The review may have conferred the most

DOI: 10.1002/acp

memory benefit to session 1 items for two reasons. First, because of the spacing effect: The size of the spaced interval between study and review was largest for the items studied in session 1. Second, the review eliminated differences between items in terms of lag from study to test, which were greatest for items studied in session 1.

Combining all 70 participants across the three experiments, spacing was more effective than massing for 63 participants (90%); three participants did equally well in the two conditions (4%); and four participants learned more in the massed condition than the spaced condition (6%). These findings are a testament to the effectiveness of spaced study using large stacks of flashcards. In their estimates of their performance at the end of session 1, by contrast, of the 58 participants who estimated that they had learned more in one condition than the other, 42 (72%) believed they had learned more in the massed condition than the spaced condition.

A final analysis concerned the time of day when participants chose to participate in the experiment. A visit to any college library reveals that college students tend to study after the sun goes down. The studying that occurred in the present experiments was no different. The percentage of sessions completed between 2 am–10 am, 10 am–6 pm and 6 pm–2 am, were, respectively, 18%, 30% and 51% in Experiment 2, and 11%, 51% and 38% in Experiment 2. Combined, 14% of sessions were completed between 2 am and 10 am, 42% were completed between 10 am and 6 pm and 44% were completed between 6 pm and 2 am.

Discussion

The addition of a review session made Experiment 3 more realistic than the first two experiments, given that virtually all students review before they take a test. Using a large stack of flashcards over multiple days (i.e. spacing) resulted in more learning than using small stacks of flashcards (i.e. massing), even in the presence of a review session.

I predicted that the review session—which amounted to a spaced study trial for the massed items, and equalised the lag to test for the spaced and massed conditions—would impact the massed condition more than the spaced condition. Contrary to this expectation, the magnitude of the spacing effect was approximately the same in Experiments 2 and 3. This finding suggests that whether or not one returns to review massed items before a test, spaced study is more effective than massed study.

Spacing was more effective than massing, but like the previous experiments, most participants believed the opposite at the end of session 1. Participants believed the two conditions had been equally effective after each of the three subsequent study sessions. However, during the review session, when participants had a chance to test themselves on both massed and spaced items after a delay, they rated spacing as more effective than massing.

GENERAL DISCUSSION

The experiments reported here represent a comparison of two common study strategies: Spacing study by repeatedly returning to a large stack of flashcards, and massing study by breaking flashcards into smaller stacks and studying them one stack at a time. The spacing strategy resulted in significantly more learning than the massing strategy in all three experiments. Combining the three experiments, 90% of participants learned more in the spaced conditions than the massed conditions, whereas only 6% of participants showed the reverse pattern.

Experiment 1 demonstrated the benefits of within-session spacing: Using a large stack of flashcards was more effective than using four smaller stacks in a single session (between-session spacing was not manipulated in Experiment 1). Experiment 2 revealed that repeating a large stack on multiple days was more effective than using a different small stack every day, and it was also more effective than massing study during the final study session, which is analogous to cramming. Experiment 3 showed that spacing was more effective than massing even if participants completed a review session, in which they were allowed to review all of the to-be-learned information, before taking the test—and surprisingly, the review session benefited the massed and spaced conditions equally. In contrast to their actual learning rates, in all three experiments, participants believed, at the end of the first study session, that massed study had been more effective than spaced study. However, with more experience, participants began to recognise the benefits of spacing.

The amount of time participants spent studying decreased dramatically between the first and last study trial on a given item. The spaced and massed conditions did not differ significantly in any experiment in terms of the amount of time participants chose to spend studying. One explanation of the spacing effect is that people pay less attention to repetitions under massed conditions than they do under spaced conditions, at least in part because learners perceive repeated items as more fluent and well learned when they are massed than when they are spaced (e.g. Dellarosa & Bourne, 1985; Greene, 1989, 1990; Metcalfe & Kornell, 2003). Based on that theory, one might predict that when learners were allowed to control their own study time, they would spend more time studying in the spaced condition than the massed condition. Previous findings are consistent with that hypothesis; for example, participants asked to read text passages twice spent less time reading the second of two massed presentations than they spent reading the second of two spaced presentations (Krug, Davis, & Glover, 1990; see also Shaughnessy, Zimmerman, & Underwood, 1927). It is unclear why participants did not spend less time studying in the massed condition than the spaced condition in the current experiments. It is possible that the massed items continued to receive study time because they were not purely massed—that is, because presentations of a given massed item were separated by study trials on other items—and because participants could test themselves as they studied.

When evaluating the effectiveness of a study technique, it is important to consider learning efficiency, as measured by the number of items learned per minute of study, as well as recall performance. A technique that increases learning 10% by doubling study time, for example, might be considered effective but not efficient. Spacing was superior to massing in terms of learning efficiency in Experiments 2 and 3. Learning efficiency may have been affected by the changes in study time across trials as well as by spacing. Spending too much time studying an item can be highly inefficient (Nelson & Leonesio, 1988). When a new word pair is presented for study, the rate of learning appears to start at a high level and quickly level off to near zero (e.g. Metcalfe & Kornell, 2003)—and the more well-learned an item is, the more quickly learning levels off. By studying quickly and speeding up over time, participants may have managed to study in a way that maintained a high level of efficiency.

Explanations of the spacing effect

Spacing study by using large stacks of flashcards had large positive effects on memory in the present experiments. There are several explanations of why spaced study is an effective way to learn. According to one class of theories, spacing is beneficial because of contextual

DOI: 10.1002/acp

variability: Spacing naturally leads to information being encoded in diverse temporal, physical and/or mental contexts, whereas massing results in the encoding taking place in a relatively fixed context (e.g. Estes, 1955; Glenberg, 1979). The spacing effect occurs because encoding information in a diverse set of circumstances results in a relatively rich, diverse set of encoding processes. In addition, the context at the time of retrieval is likely to share more cues with a diverse set of stored cues than with a homogeneous set of stored cues. Compared to standard laboratory experiments on spacing, the spaced conditions Experiments 2 and 3 involved a high degree of contextual fluctuation: Not only did the sessions take place on different days, but they took place at different times of day, and potentially in different places, on different computers, while listening to a variety of music, with different friends in the room, and so on. The massed sessions each took place in only one of those environments; they did not span different environments (except in Experiment 3, when the massed items were studied in the review session). Thus, in this and other online studies of spacing (Cepeda et al., 2006), contextual variability may have been a relatively important cause of the spacing effect.

Another explanation of the spacing effect is diminished effort in the massed condition: People sometimes pay less attention to the second presentation of a massed item than they pay to the second presentation of a spaced item, because the massed item is already highly familiar the second time it is presented (e.g. Dellarosa & Bourne, 1985; Greene, 1989). For example, measurements of pupil dilation indicate that people pay more attention on the second of two spaced learning opportunities than they do on the second of two massed learning opportunities (Magliero, 1983). Also, people are more accurate in a secondary detection task when they are studying the second presentation of a massed item than when they are studying the second presentation of a spaced item (i.e. they pay more attention to the spaced item; Johnston & Uhl, 1976). In part, people may make less effort because they believe they have already learned the massed items better than they have learned the spaced items (Zechmeister & Shaughnessy, 1980). In the present experiments, the fact that four cues and four targets were presented between repetitions of a given item in the massed condition was probably enough to clear a given item from working memory before that item recurred. As a result, recalling the answer on the back of the card required effort even in the massed condition (although more effort was presumably required in the spaced condition). Thus diminished effort probably played a relatively small role in the spacing effect in the present experiments. The study time data are consistent with this hypothesis: In a self paced study situation, diminished effort should translate into participants spending less time studying massed items than spaced items (Krug et al., 1990; Shaughnessy et al., 1972). There was no such study time difference in the present experiments.

A third way of understanding spacing effects is based on accessibility (Bjork & Bjork, 1992). Learning appears to be a function of a memory's current level of accessibility when it is relearned: The less accessible a memory is, the more learning occurs when it is restudied (Bjork & Allen, 1970; Cuddy & Jacoby, 1982; Krug et al., 1990). Because spacing allows time for items to be forgotten between study opportunities, spaced items are, on average, less accessible the second time they are studied than are massed items. This reduced accessibility may enhance learning in the spaced condition (if the initial memory can be retrieved; see Appleton-Knapp, Bjork, & Wickens, 2005). Accessibility may be especially important when, as in the present experiments, participants attempt to recall an item before its answer is presented. In the present experiments, spacing reduced accessibility in two ways, by increasing the number of items in-between successive study opportunities on a given item, and, in Experiments 2 and 3, by allowing entire days to pass between study opportunities, drastically increasing forgetting. The reduced accessibility of the spaced items during study may also have contributed to participants' belief that spacing study was less effective than massing study.

The accessibility account can also explain the difference, in Experiment 3, between items presented in the spaced condition and items presented during session 1 in the massed condition and then reviewed during session 5. The total spacing of the two types of items was the same, and yet the spaced items were recalled at a higher rate than the massed items. According to the accessibility account, the spaced condition repeatedly decreased the accessibility of the forgotten items to low levels at the beginning of each study session, making each session an important study opportunity. In the massed condition, accessibility was maintained at a high level during session 1, and only decreased significantly one time, between session 1 and session 5.

Why learners prefer massing over spacing

At the end of session 1, participants in all three experiments believed that massing had been more effective than spacing. In Experiment 3, for example, participants predicted that final test accuracy—which was 31 percentage points higher in the spaced condition than the massed condition—would be 14 percentage points *lower* in the spaced condition than the massed condition. Presumably, participants tested themselves as they studied, and those tests were more successful in the massed condition (with only four items intervening between repetitions) than in the spaced condition (with 19 items intervening between repetitions), and moreover, the massed items were presented for study eight times during the session 1, whereas the spaced items were presented only twice. Thus at the end of session 1, the massed items were more accessible and fluently retrievable than the spaced items. Participants may have assumed that short-term retrieval fluency signified long-term learning, which is an all-too-common mistake (Benjamin, Bjork, & Schwartz, 1998; Kelley & Lindsay, 1993). Maximizing long-term learning requires creating challenges for learners that engage the type of active information processing that creates lasting memories (Bjork, 1994, 1999). During sessions 2, 3 and 4, participants estimated that they had learned the same amount in the spaced and massed conditions. Given that recall was higher in the spaced condition than the massed condition in all three experiments, these estimates, although not as egregious as the estimates made during session 1, were nonetheless inaccurate. Even after the final test, in Experiment 1, participants did not give higher ratings to spaced than massed study.

The impression participants formed at the end of the first session is probably what matters most from a practical perspective. Anecdotal evidence suggests that once students find a study technique that they think works, they rarely choose to experiment with other study techniques. Those students who preferred massing after session 1 would probably have abandoned spacing and decided to mass their study from that point forward.

For a learner to accurately assess the effectiveness of spaced vs. massed study using flashcards becomes even more complex when learners change the size of their stack as they study. Learners frequently remove items that they feel they have already learned from a stack of flashcards, which can, unfortunately, have negative effects on learning—perhaps in part because doing so decreases the size of a stack of flashcards, but also in part because learners often remove items that they do not yet know (Kornell & Bjork, 2008b). Cards can be added to a stack as well, and some authors have suggested that, as learning progresses on

DOI: 10.1002/acp

a given item, the optimal study schedule may involve gradually increasing the intervals between study trials on that item, while simultaneously introducing new items over time (Landauer & Bjork, 1978; Mondria & Mondria-De Vries, 1994).

CONCLUSION

The present findings suggest some clear practical advice for students: To be efficient, flashcards should be studied in relatively large stacks across multiple days. Moreover, spacing is more effective than cramming, even if total study time is controlled. Furthermore, learners who perceive massed study as more effective than spaced study should beware: Massed study is seductive, and it can appear to be more effective than spaced study even when spaced study is the more effective strategy.

The present experiments mimicked real flashcard study in a number of ways. The massed condition was not purely massed (i.e. participants never studied the same flashcard twice in a row); the timing of presentations was self-paced and participants were not required to make overt responses while studying; the materials were GRE-type words; participants studied on their own time, often in the middle of the night, and in their own study environment (e.g. the campus library, a dorm room, on the couch at home); and there was a review session before the test in Experiment 3. These aspects of the experiments demonstrate that the spacing effect can be generalised to a real-life study situation. The present findings may extend beyond flashcards, as well; for example, relatively short practice sessions distributed evenly across days may be more effective than intense but infrequent practice sessions for musicians, athletes, pilots and learners in a wide variety of other domains.

ACKNOWLEDGEMENTS

I thank R. Blaize Wallace for his help in designing and programming the experiment and Robert A. Bjork for his advice and support. Grant 29192G from the McDonnell Foundation supported this research.

REFERENCES

Appleton-Knapp, S., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes and their interaction. *Journal of Consumer Research*, *32*, 266–276.

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21*, 627–635.

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

Benjamin, A. S., & Bird, R. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory & Language*, *55*, 126–137.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*, 567–572.

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

Brinthaupt, T. M., & Shin, C. M. (2001). The relationship of academic cramming to flow experience. *College Student Journal*, *35*, 457–471.

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.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*, 451–467.

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory & Cognition*, *13*, 529–537.

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. (1996). Distributing and managing the conditions of encoding and practice. In R. Bjork, & E. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

Dunlosky, J., & Nelson, T. O. (1992). Importance of kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, *33*, 545–565.

Ebbinghaus, H. E. (1964). *Memory: A contribution to experimental psychology* (Henry A. Ruger & Clara E. Bussenius, Trans.). New York: Dover. (Original work was published in1885).

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

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*, 19–34.

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

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 371–377.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1004–1011.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–97). Potomac, MD: Erlbaum.

Johnston, W. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning & Memory*, *2*, 153–160.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224.

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the 'enemy of induction'? *Psychological Science*, *19*, 585–592.

Kornell, N., & Bjork, R. A. (2008b). Optimizing self-regulated study: The benefits-and costs-of dropping flashcards. *Memory*, *16*, 125–136.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 609–622.

Krug, D., Davis, B., & Glover, J. A. (1990). Massed versus distributed repeated reading: A case of forgetting helping recall? *Journal of Educational Psychology*, *82*, 366–371.

DOI: 10.1002/acp

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Magliero, A. (1983). Pupil dilations following pairs of identical and related to-be-remembered words. *Memory & Cognition*, *11*, 609–615.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

Mondria, J. A., & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. *System*, *22*, 47–57.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207–213.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the 'labor-in-vain effect'. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 676–686.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions In Psychological Science*, *16*, 183–186.

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

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems improves learning. *Instructional Science*, *35*, 481–498.

Salisbury, D. F., & Klein, J. D. (1988). A comparison of a microcomputer progressive state drill and flashcards for learning paired associates. *Journal of Computer-Based Instruction*, *15*, 136–143.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107–122.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *11*, 1–12.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 907–912.

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction*, *1*, 341–358.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 601–604.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky, & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.

Spellman, B. A., & Bjork, R. A. (1992). Technical commentary: When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*, 315–316.

Vacha, E. F., & McBride, M. J. (1993). Cramming: A barrier to student success, a way to beat the system, or an effective learning strategy. *College Student Journal*, *27*, 2–11.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, *15*, 41–44.

APPENDIX

Word pairs used in the present experiments.

Cue	Target
Abrogate	Abolish
Apotheosis	Deification
Bluster	Swagger
Chimerical	Fantastical
Collusion	Conspiracy
Commodious	Ample
Complaisant	Accommodating
Connubial	Marital
Convivial	Festive
Coquette	Vixen
Dereliction	Abandonment
Descry	Detect
Desultory	Aimless
Dilatory	Delaying
Doctrinaire	Inflexible
Effulgent	Brilliant
Encomium	Praise
Enervate	Weaken
Equipoise	Equilibrium
Exiguous	Miniature
Feckless	Ineffective
Fulsome	Sickening
Importunate	Demanding
Insuperable	Undefeatable
Interstice	Aperture
Limpid	Serene
Manumit	Emancipate
Mottle	Blotchy
Peccadillo	Misdeed
Quotidian	Commonplace
Rabble	Mob
Rectitude	Righteousness
Reticulose	Network
Ribald	Vulgar
Specious	Illogical
Stygian	Hellish
Supine	Passive
Sycophant	Flatterer
Torpid	Sluggish
Voluble	Talkative