APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **14**: 215–235 (2000)

Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall

WILLIAM L. CULL*

Loyola University of Chicago, USA

SUMMARY

Spacing multiple study opportunities apart from one another is known by psychologists to be a highly effective study method (see Dempster, 1996). This study examines whether including tests during study would produce practical benefits for learning beyond that provided by distributed study alone. In addition, spacing of both study and test (massed, uniform distributed, and expanding distributed) is investigated. To-be-remembered information was repeated with a single learning session (Experiment 1), reviewed immediately after initial learning (Experiment 2), or reviewed days after initial learning (Experiments 3 and 4). As expected, large distributed practice effects were shown across experiments. In addition to these effects, testing produced significant benefits for learning in all four experiments, which were of moderate or large size (Cohen's *d* of 0.52 to 1.30) for three experiments. Expanding test spacing, however, did not independently benefit learning in any of the learning situations studied. Educators should take advantage of the large benefits that distributed study and testing have on learning by spacing multiple tests of information within learning sessions and by distributing tests across multiple review sessions. Copyright © 2000 John Wiley & Sons, Ltd.

The psychological study of human memory and learning has yielded many principles that educators may use to improve student learning. Three principles regarding the best methods to restudy information are considered in the current study. The first principle, based on the frequently investigated spacing effect, is that study opportunities distributed across time are more effective in promoting learning than are study opportunities that are unspaced or massed together (Hintzman, 1974; Melton, 1970, Underwood, 1970). The second principle, sometimes referred to as the retrieval effect, is that testing a learner's knowledge not only provides an indication of what is known but also improves the learner's understanding of that information (Carrier and Pashler, 1992; Dempster, 1996). The third principle, known as the expanding-test-series mnemonic, is that when multiple tests are given to learners, a pattern of tests that has a short time interval before the first test and has increasingly larger time gaps between subsequent tests is most effective (Cull *et al.*, 1996; Landauer and Bjork, 1978; Rea and Modigliani, 1985).

*Correspondence to: William L. Cull, Department of Medical Education (MC 591), The University of Illinois at Chicago, 808 S. Wood Street, 976-M CME, Chicago, IL 60612-7309, USA. Tel: (312) 996-7653
E-mail: WCull@UIC.edu

Received 25 January 1998
Accepted 28 January 1999

Practical benefits for education have been suggested for distributed study (Dempster, 1988, 1996), testing (Dempster, 1992; Dempster and Perkins, 1993), and expanding test spacing (Baddeley, 1990; Bjork, 1994; Schmidt and Bjork, 1992). It has also been suggested that all three principles should be incorporated simultaneously in learning programs in order to maximize performance (Dempster, 1996). The purpose of this study is to examine whether these effects that are closely tied to one another and have often been entangled in the research literature can bestow meaningful benefits for learning that are independent of each other.

Although research confirms that testing improves learning, it is not clear that testing provides educationally important benefits beyond those provided by distributed practice. Several of the first studies on testing demonstrated that providing tests that were embedded within an initial learning session (test repetitions) or providing tests across separate learning sessions (test reviews) produce greater learning than when no tests are provided (Bartlett, 1977; Bartlett and Tulving, 1974; Darley and Murdock, 1971; Madigan and McCabe, 1971; McDaniel *et al*., 1989; Modigliani, 1976; Runquist, 1983, 1986a, b, Young 1971). These studies, however, did not rule out the possibility that the tests merely provided the learner with an additional learning opportunity; thus, it was not clear that the attempt to retrieve information from memory itself benefited learning. Other research comparing testing with comparable amounts of additional study have showed an advantage for testing (Allen *et al*., 1969; Landauer and Bjork, 1978; LaPorte and Voss, 1975; Hogan and Kintsch, 1971; Wenger *et al*., 1980), while others either showed an advantage for additional study (Landauer, 1969; McDaniel and Masson, 1985) or no difference between additional study and testing (Donaldson, 1971; Landauer and Eldridge, 1967; Nungester and Duchastel, 1982; Whitten and Bjork, 1977).

More recently, Carrier and Pashler (1992) argued that comparisons of testing and additional study conditions were biased toward the study condition. Specifically, they argued that because no feedback concerning correct answers was provided in the test condition, restudy was contingent upon the learner being able to recall successfully the information from memory. In the study condition the learner was always provided with complete item information, ensuring that the learner could restudy the to-be-learned information. Thus, tests may promote more effective processing than mere study when information is remembered, but this effect may be overshadowed by the inability of learners to remember many items.

Carrier and Pashler provided a solution to this problem by adding a feedback component to the test condition while keeping the total study time constant across conditions. When they presented learners with either a study presentation and two test-study repetitions or with three study presentations, recall was significantly higher for the test-study condition on final tests given either 5 minutes or 24 hours later. These findings suggest that testing does have an effect on learning that is independent of providing additional study, at least when information is tested within a single study session. It remains untested whether similar benefits hold when tests are provided in separate learning sessions (i.e. review tests) or when other practice schedules are used. For example, would a student preparing for an exam across multiple days also be better served to have reviews with tests rather than reviews with additional study?

The amount of research investigating the use of expanded-spacing practice schedules is much less than that for distributed study or testing, but the results have been promising. Landauer and Bjork (1978) first demonstrated that expanded-spacing

practice schedules are more effective for learning names than are uniform-spaced schedules, contracting-spaced schedules, massed schedules, or not providing tests. This was shown for learning tasks where learners were asked to pair first names with last names or to pair names with faces. Recently, this effect was replicated when name pairs were presented visually or auditorially and when trivia facts were learned using experimenter-paced or learner-paced study procedures (Cull *et al*., 1996). Expanding test spacings were found to be especially helpful for poor learners. However, when feedback was provided following the tests, expanding spacings were not shown to be superior to uniform distributed spacings, although high performance levels may have influenced this finding.

Providing feedback following testing is likely to be important when students' study strategies can benefit from feedback and when knowing what is the correct answer can increase understanding. Rea and Modigliani (1985) showed that schoolchildren learned spelling and multiplication facts best when an expanding schedule of tests with feedback (test-study trials) was used. Unfortunately, they did not compare the expanding schedule with a uniform distributed schedule of tests to separate the benefits of expanded testing from the benefits of distributed practice. More work needs to be done to see if expanding-testing effects are present for tests with feedback. Moreover, research to this point has focused exclusively on repetition tests within a learning session. Although some researchers suggest that expanded spacing should also be considered when tests are given across multiple learning sessions that are spaced days apart (Bjork, 1994; Dempster, 1996; Sones and Stroud, 1940), no empirical studies have demonstrated a benefit of expanded spacing for those review tests.

The specific aims of this study are (1) to compare the effect of testing and expanding test spacing when tests with and without feedback are provided within a single learning session (Experiment 1), immediately after initial learning (Experiment 2), and days after initial learning (Experiments 3 and 4), (2) to untangle the effects of distributed study, testing, and expanding test spacing, by assessing whether testing and expanding spacing produce benefits beyond that provided by distributed study, and (3) to indicate the size of these effects in order to help evaluate their applied educational value.

EXPERIMENT 1

In Experiment 1 participants were asked to learn a pool of discrete verbal items within a single learning session that was fairly brief. This learning situation is similar to those where distributed study, testing, and expanding test spacing effects have been most frequently studied, and it is similar to many educational situations. For example, students often are asked to use a flash card type procedure to learn a list of information such as multiplication facts, spelling words, or foreign vocabulary words. The present experiment investigated what spacing schedules of repetitions are most effective and whether test repetitions are more effective than study repetitions. Four spacing schedules were manipulated in this experiment: an expanding distributed spacing, a uniform distributed spacing, a massed spacing, and a nonrepeated control group. Also, three types of presentations were manipulated: study-only repetitions, test-only repetitions, and test-study repetitions. If distributed practice, testing, and expanding test spacing all have positive effects on learning, then the combination of

expanding distributed spacing and test-study presentations should produce higher performance than all other treatment combinations.

Method

Participants
A total of 66 students from the introductory psychology subject pool at Loyola University of Chicago participated in the experiment.

Materials
All participants were asked to learn, using an IBM-compatible computer, a list of 32 paired-associate items of moderate difficulty according to Underwood's (1982) norms. An additional 12 items, also of moderate difficulty, were used as filler items. Only paired-associates that consisted of an uncommon word as the cue member and a common word as the response member were selected. This made the word pairs similar to unfamiliar vocabulary wherein the to-be-learned word is uncommon, and the definition consists of common words. Both words of each pair were five letters in length. An example pair is *bairn-print*.

Procedure
Participants were randomly assigned to study the items in one of three ways. In all three conditions, participants were presented the same items for study on four different occasions; the first presentation was always a study presentation with both the cue and response word presented, one above the other, for the duration of the presentation. In the first condition (test-study), participants were presented the cue alone for 6 seconds followed by the cue and response for 2 seconds on the remaining three presentations. In the second condition (study-only), the cue and response were presented for 8 seconds on each of the remaining three presentations. Finally, in the third condition (test-only), the cue member alone was presented for the entire 8-second period on the remaining presentations.

In the test-study and test-only conditions, a cursor box was presented beneath the cue word when the cue was presented alone. Learners were asked to type in the response word if they could remember it and were informed that it would be difficult to enter a response successfully within the time period allowed. They were also assured that they would have unlimited time to enter their response on the final recall test.

In addition to the between-subjects variable, type of presentation, the spacing of presentations and the duration of the initial study presentation were also manipulated as within-subjects variables. Participants were presented all 32 critical items according to their respective type of presentation (study-only, test-only, or test-study), but the items were randomly divided into four subsets (8 items) that were randomly assigned to each of four spacing schedules (expanding, uniform, massed, or non-repeated). In the massed condition (0–0–0 items), all four presentations or tests of an item occurred consecutively; there were no intervening items. In the uniform condition (5–5–5 items) each repetition of an item occurred after a uniform gap of five other items. In the expanding condition (1–5–9 items), repetitions were also distributed but based on increasing intervals between items: a 1-item gap separated the first and second presentations of the item, a 5-item gap separated the second and third presentations, and a 9-item gap separated the third and fourth presentations.

Both critical and filler items were used to create the gaps between item repetitions. Non-repeated items served as control items and were only presented one time with both cue and response presented.

Half of the eight items assigned to each spacing condition were then randomly selected to have shorter durations on their first or initial presentations. Items in the short condition were presented for 4 seconds on their first presentation; items in the long condition were presented for 8 seconds. Short and long items only differed for the first presentation. (This was done to vary degree of learning prior to restudy.) All subsequent presentations remained 8 seconds in length. For each participant, the items were assigned randomly to conditions and were ordered for presentation within a condition randomly. Thus, the items that were assigned to each of the condition combinations and the ordering of items differed for each of the learners.

Once all the items had been studied, participants were given a 1-minute filler task (they were asked to read instructions for the retention test) followed by a computerized cued-recall test of all 32 items. On the cued-recall test, the cue word from each of the originally studied items was presented along with a cursor box and a prompt that asked participants to type in the appropriate response word. Participants were given as long as they needed to enter a response on the final test, but they were not able to return to an item once a response had been entered or the item had been skipped. The order of the items again was random and thus different for each participant.

Results and discussion

Across all experiments, an alpha-level of $p \leqslant 0.05$ was used for all statistical tests conducted. The proportions of correct responses on the final cued-recall test of Experiment 1 are listed in the top portion of Table 1. A $3 \times 4 \times 2$ mixed ANOVA was used to examine the effects of type of repetition, spacing of repetition, and length

Table 1. Final test proportion recall as a function of type and spacing of restudy

	Type of restudy				Effect size (Cohen's *d*)
Spacing	Test-study	Study-only	Test-only	Non-reviewed	Test-study versus study-only
Experiment 1 – Repetitions					
Uniform Distributed	0.49	0.29	0.34	–	0.66
Expanding Distributed	0.48	0.27	0.38	–	0.71
Massed	0.19	0.14	0.18	–	0.24
Experiment 2 – Immediate reviews					
Uniform Distributed	0.78	0.65	0.35	0.34	0.52
Expanding Distributed	0.70	0.56	0.34	0.26	0.57
Massed	0.72	0.62	0.32	0.22	0.41
Experiment 3 – Delayed reviews (3-day retention interval)					
Uniform Distributed	0.98	0.91	0.64	0.27	0.33
Expanding Distributed	0.84	0.78	0.55	0.18	0.28
Massed	0.49	0.34	0.50	0.14	0.70
Experiment 4 – Delayed reviews (8-day retention interval)					
Uniform Distributed	0.89	0.64	0.69	0.20	1.30
Expanding Distributed	0.82	0.60	0.51	0.19	1.14
Massed	0.42	0.21	0.35	0.20	1.09

of initial presentation on recall. Length of initial presentation was not found to have a significant effect on recall, or to interact with either of the other variables. Thus, the results, as summarized in Table 1, were combined for short and long initial presentations. A marginally significant main effect was found for type of repetition, $F(2,63) = 2.57$, $MSe = 4.30$, $p = 0.09$, and a significant main effect was found for spacing, $F(3,189) = 68.11$, $MSe = 0.65$. These effects were influenced, however, by a significant interaction between these variables, $F(6,189) = 2.66$, $MSe = 0.65$.

A detailed investigation of the interaction revealed that all spacings of repetitions produced significantly higher recall than did the non-repeated control group whose mean proportion recall was 0.09 (test-study, $F(1,21) = 4.34$, $MSe = 1.03$; test-only, $F(1,21) = 9.27$, $MSe = 0.63$; study-only, $F(1,21) = 8.10$, $MSe = 0.55$). Spacing effects were also found for all three types of repetitions; as recall for the distributed spacing conditions was significantly higher than recall for the massed spacing condition (test-study, $F(1,21) = 43.73$, $MSe = 1.35$; test-only, $F(1,21) = 17.46$, $MSe = 0.95$; study-only, $F(1,21) = 6.79$, $MSe = 1.62$). The interaction of type and spacing of repetitions resulted from greater differences between the distributed and massed spacings for the test-study and test-only conditions rather than for the study-only condition. Unexpectedly, the expanding distributed spacing condition did not produce significantly higher recall than the uniform distributed spacing condition for any type of repetition.

Significant testing effects were found in addition to the distributed practice spacing effect. Participants in the test-study condition recalled reliably more than did participants in the study-only condition for expanding items, $F(1,63) = 7.03$, $MSe = 4.43$, and for uniform items, $F(1,63) = 5.56$, $MSe = 5.01$, but not for massed items. The first panel of Figure 1 further shows the benefit of testing in comparison to distributed study and to the non-repeated control group. These effects were combined for expanding and uniform spacing directions, and they show sizable independent benefits of distributed study and testing. Table 1 shows that participants in the test-only condition also recalled more than study-only participants did, but these differences were not significant.

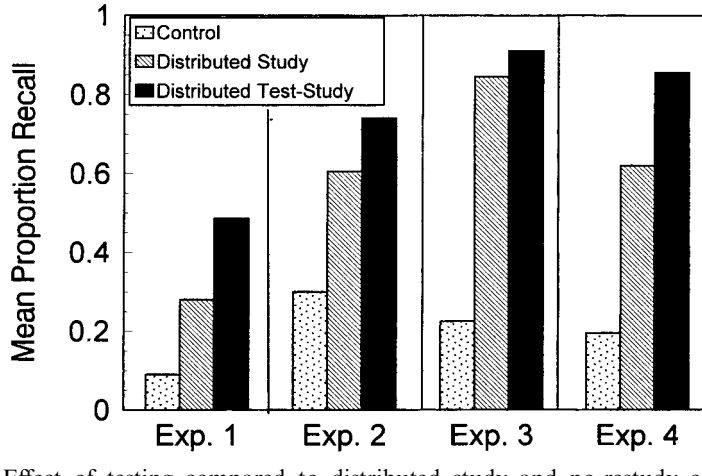

Figure 1. Effect of testing compared to distributed study and no-restudy control groups across experiments

Cohen's *d* was used as a measure of effect size to represent the difference between the test-study and study-only conditions (Cohen, 1992). These values are presented on the right-hand side of Table 1. Cohen (1992) categorized small, medium, and large effect sizes as 0.20, 0.50, and 0.80, respectively. Using this scale, medium effects of testing were found for both expanding and uniform distributed spacing schedules. These benefits of testing are consistent with those demonstrated by Carrier and Pashler (1992).

From a theoretical perspective, several findings from Experiment 1 are interesting. First, expanding spacings were not found to be significantly more effective than all other spacing schedules. This finding is inconsistent with previous research showing expanding test spacing effects (Cull *et al*., 1996; Landauer and Bjork, 1978), but it should be noted that the present study means were in the expected direction for test-only repetitions, the condition where this effect has been previously demonstrated. Second, a significant testing effect was not found for massed items. This indicates that the testing effect does not generalize across all spacing schedules. Finally, the length of the initial item presentations, which varied from 4 to 8 seconds did not affect final recall performance within this multiple-presentation learning situation. Apparently, the extra time for the initial presentation in the long condition represented wasted time for learners or labor in vain (see Nelson and Leonesio, 1988), or perhaps the obscurity of the cue words made more elaborate encoding of information difficult upon initial presentation, minimizing the potential advantage of the longer initial presentation.

EXPERIMENT 2

Experiment 2 examined the potential benefits of distributed practice, testing, and expanding test spacing in a learning situation where reviews of information were provided just after initial study. In the initial-learning phase of Experiment 2, each learner cycled through the study items at his or her own pace and was allowed to study each item multiple times. Repetitions of items were spaced in order to take advantage of the type of distributed practice that was shown to be effective in Experiment 1. Then, following a clear break that indicated that initial learning had been completed, participants were provided with multiple review sessions. The time between initial learning and these reviews were very brief (1–10 minutes), and all learning, reviewing, and final testing took place within a 1-hour time period. The same spacing patterns that were used in Experiment 1 were used in Experiment 2, but these spacings reflected the time between review sessions as opposed to time between item repetitions. This learning situation is similar to that experienced by a student who, just after reading a textbook chapter, answers the review questions given at the end of the chapter. These reviews are designed to help learners consolidate their understanding of the information that had just been studied. Of primary interest in this experiment was whether review in the form of tests would be more effective than review involving additional study in improving learning and whether certain spacings of review tests would be more effective.

Method

Participants

Sixty-six students participated from the same subject pool that was used for Experiment 1. No participant had been a participant in the previous experiment.

Materials and procedure

The list of to-be-learned items was the same as the list used in Experiment 1. The experiment was administered using an IBM-compatible computer. Again, two variables were manipulated: the type of review (study-only, test-only, and test-study) and the spacing of review (expanding, uniform, and massed). In contrast to Experiment 1, the type of review was manipulated within subjects and the spacing of review was manipulated between subjects.

Participants were assigned randomly to one of three spacing conditions and were first asked to complete the initial learning phase, which was identical for all three conditions. During this initial learning period, all 36 words were available for study within a 400-second period (6 minutes 50 seconds). An item was presented on the screen with the cue above the response and a prompt in the lower left-hand corner reading 'Press SPACEBAR to see the next item or "d" to drop that item'. That pair remained on the screen until the space bar or letter 'd' was pressed. If the spacebar was pressed, another item was immediately presented on the screen along with the prompt, but the item remained in the list of to-be-learned items. If 'd' was pressed, then that item was removed from the list of items to-be-learned and the next item was presented. The items were presented in cycles such that all items were presented before any item was re-presented. For each cycle, the items were presented in a random order and that random order was different for each participant. The initial study session was designed to be like studying from a set of flash cards since the participants were able to circulate through the items as quickly or slowly as they desired and to discard items that had been learned in order to concentrate on the remaining items.

Once the initial study phase was over, participants were given a series of three reviews that were separated from one another by having participants perform a vocabulary rating distracter task. The three spacing conditions differed in terms of when the reviews were provided relative to the rating of 14, 14-item distracter lists. Each distracter list took roughly 2 to 3 minutes to complete. (The administration of distracter lists and reviews is described below.) In the expanding condition (0–3–6 lists), the first review session was provided immediately after initial study; the second review session was provided after three distracter lists had been completed; and the third review session was provided after another six distracter lists had been completed. In the uniform condition (3–3–3 lists), the initial study and review sessions were separated by three gaps of three distracter lists, and in the massed condition (0–0–0 lists), the initial study session and review sessions were presented in succession prior to the administration of all 14 distracter lists. For the expanding and uniform conditions, the remaining five distracter lists were inserted as a filler between the last review session and the final retention test.

The 14 distracter lists were formed from a random sample of words that were taken from the *Oxford American Dictionary* (see D'Anna *et al.*, 1991). Each distracter list consisted of 14 items. For each 14-word list, the words were randomly presented one at a time. Each word first appeared alone for 1.5-seconds followed by the prompt: 'Please rate your knowledge of this word.' A scale also appeared: '1 – I have never seen this word before, 2 – I have seen this word but do not know its definition, 3 – I have seen this word and have some idea of its meaning, 4 – I know this word well enough to recognize its meaning, or 5 – I know this word well enough to define it.' The word and the scale appeared on the screen for 15-seconds or until a rating was made,

whereupon the next item was presented. There was a distinct beginning and end to each distracter list.

For each review session, the critical items were divided randomly into four sets of eight items and assigned to one of four conditions: study-only, test-only, test-study, or no review. In the study-only condition, the cue as well as the response were presented to the participant for a 12-second period. In the test-only condition, the cue was presented without the response for the 12-second period; a cursor box also was presented, and participants were prompted to attempt to type the appropriate response and press return. If an answer was registered before the 12-second period was over, a message appeared letting the participant know that the answer had been registered, but the next item was not presented until the entire 12-second period had expired. In the test-study condition, the cue was presented alone with a cursor box for the first 8-seconds wherein participants were asked to enter an appropriate response; this was followed by a 4-second period where the cue and response were presented together. Again, if an answer was registered prior to the completion of the initial 8-second test period, a message verified the registration, but the study period did not begin until the entire 8-second test period was over. Control items (no review) were not presented in any of the review sessions.

Items remained in their respective conditions (study-only, etc.) for all review sessions. Within a review session, the order of item presentation was completely random for each participant, and the order was randomly shuffled before each review session. Once all review sessions and distracter lists had been administered, the final recall test was administered just as in Experiment 1.

Results and discussion

The results for the final cued-recall test of Experiment 2 were analysed using a 4×3 mixed ANOVA design. The proportions of correct responses are summarized in Table 1. The results showed a significant main effect for the type of review, but no significant main effect of spacing or interaction between spacing and type of review. A detailed analysis of the main effect of type of review revealed that all three types of immediate review, test-study review, study-only review, and test-only review, significantly improved performance in comparison to the non-reviewed control group, $F(1,65) = 5.05$, $MSe = 1.64$.

As was the case in Experiment 1, testing was shown to improve learning performance significantly. The test-study condition produced much greater recall than the study-only condition, $F(1,65) = 31.09$, $MSe = 1.06$. Table 1 shows that the effect sizes[1] were of moderate size for the expanding ($d = 0.57$) and uniform ($d = 0.52$) distributed spacings and small to moderate in size for the massed spacing ($d = 0.41$). The second panel of Figure 1 further shows the effect of testing in comparison to the effect of review. The effects of testing were again combined for the expanding and

[1]The computation of the effect size measure, Cohen's *d*, is computed for between-subjects comparisons by dividing the difference between the two treatment means being compared by the within-population standard deviation (Cohen, 1992). An approximation of this measure was used in Experiments 2–4, because the conditions being compared were manipulated within subjects, making values within the distribution related. The standard deviation used to compute Cohen's *d* was determined by averaging the number of items recalled within the two groups being compared, and then pooling the standard deviations of those averages across the between-subjects conditions.

uniform distributed spacings because no advantage was found for an expanding spacing of immediate reviews.

The results of Experiments 1 and 2 revealed two primary differences. First, no difference between the massed spacing condition and the distributed spacing conditions was found in Experiment 2. It appears that the spacings, which differed from each other in terms of minutes of distracter activity, were not sufficiently different from each other in order to produce the normally robust spacing effect. Second, test-only items were recalled significantly less than study-only items were, $F(1,65) = 64.66$, $MSe = 2.43$. This is likely related both to a reduced likelihood of successful recall on immediate review tests in Experiment 2 rather than on repetition tests in Experiment 1 and to greater effectiveness of the study-only condition in Experiment 2.

EXPERIMENT 3

Experiment 3 investigated whether the beneficial effects of distributed practice, testing, and expanding testing would be found in a more educationally relevant learning situation that used actual vocabulary words and spaced multiple review sessions across days. The reviews were spaced across 6 days, and there was a 3-day retention interval between the last review session and the final retention test. Having multiple learning sessions spaced across days is a learning situation that students experience often when preparing for quizzes and examinations. The initial learning session utilized distributed test-study presentations. It was of interest whether the inclusion of testing in these reviews also would be beneficial and whether certain review spacings would be more effective for learning than others.

Method

Participants

A total of 54 students from the subject pool used in the previous experiments and from two upper-division psychology courses at Loyola University of Chicago participated. Only 43 of those who started the experiment successfully completed the experiment; four participants, three participants, and four participants, respectively, were lost from the expanding, uniform, and massed spacing conditions. No participant had participated in any of the earlier experiments.

Materials and produce

Pilot testing suggested that when learning was examined across a 9-day period, a number of changes in the stimulus items and initial learning procedures needed to be made in order to prevent extremely low recall levels. First, to involve participants in the task to a greater extent, a list of uncommon but real vocabulary items was used (McDaniel and Pressley, 1984). An example item is *handsel-payment*. Second, unlike the procedure followed in the previous experiments, only the initial learning phase was computerized in Experiment 3.

As in Experiment 2, spacing of review was manipulated as a between-subjects variable and type of review was manipulated as a within-subjects variable. Participants were randomly assigned to one of three spacing conditions. Participants in all conditions first completed an initial learning phase. In this initial learning period,

participants were first presented with 40 vocabulary words using the computerized flash-card study procedure used in Experiment 2. Unlike Experiment 2, there was no time limit for the initial study period. Study continued until all of the items had been dropped out of the to-be-studied list. Then, all the cue members of the items were presented along with a cursor box and a test prompt asking participants to type the appropriate definition or response member for each word. If participants answered correctly, the program indicated that the response was correct and the item was dropped from the list; if participants answered incorrectly, the correct response was provided and that item was placed in a set of items that would be re-presented. Once all items had been tested, those items that were answered incorrectly continued to be restudied and retested in new random orders until the participant answered each item correctly once. This took roughly 30 minutes on average to complete.

Participants within each of the spacing conditions then received review sessions and rating sessions at different spacings. Each review session was administered using a flash-card booklet that was prearranged to provide, in a random order, study-only review (one card with a word and definition) for 10 items, test-only review (one card with just a word) for 10 items, test-study review (one card with just a word followed by a second card with the same word and definition) for 10 items, and no review for a remaining 10 items. Four different booklet orders were created so that across participants each item served in each of the within-subjects conditions equally often. Participants reviewed the items at their own pace, but they were asked not to go backwards through the booklet at any time. For test items, participants were asked to write the appropriate definition on the test card. Rating sessions presented the same pool of distracter items as presented for Experiment 2, using three different rating sheets comprised of 66, 66, and 67 items respectively. Rating sessions were used in this experiment to balance the number of days that participation was required for each of the spacing conditions to prevent any selective loss of participants; if rating sheets were not used, participants in the massed condition would only have participated on 2 days and participants in the expanding and uniform conditions would have participated on 5 days.

Participants in the expanding spacing condition (1–2–3 days) were asked to complete all three rating sessions immediately after initial study. Then, following a 1-day delay, participants completed the first review booklet, followed by the second review booklet 2 days after that, and the final review booked 3 days after that. Participants were asked to complete the review booklets at any time within the designated day and that completion of the review booklets was not monitored. A final retention test that was monitored was given 3 days after the final review session. The same procedure was used for the uniform spacing condition (2–2–2 days) except that all review sessions were spaced at 2-day intervals with the final retention test given 3 days after the final review. The massed spacing condition (0–0–0 days) differed from the other two spacing conditions in that all three review sessions were administered immediately after initial study. So that massed participants would also have to complete parts of the experiment on their own, they were asked to complete the three rating sessions according to either an expanding or a uniform spacing. Expanding ratings were spaced at expanding gaps of 1, 2, and 3 days, and uniform gaps were spaced at gaps of 2 days each. Roughly, an equal number of participants within the massed condition received expanding and uniform spacings of ratings.

In all spacing conditions, participants were given the final cued-recall test 9 days after the initial study session. The final test was administered in the laboratory using a pencil and paper procedure. Cue-members for all critical items were arranged in the same random order for all participants. Participants were given as much time as they needed to write in a definition for each of the words.

Results and discussion

As previously mentioned, 11 participants did not successfully complete the experiment. These participants either failed to complete a review or rating session on the scheduled day or they failed to return for the final test. An additional participant within the uniform condition was randomly excluded from all analyses to provide an equal number of 14 participants within each of the conditions.

The results for the final cued-recall test of Experiment 3 are summarized in Table 1. A 4×3 mixed design ANOVA was used to analyse these results. A significant interaction was found between the spacing and type of review, $F(6,117) = 7.85$, $MSe = 2.88$, in addition to significant main effects for each variable: spacing, $F(2,39) = 13.41$, $MSe = 11.78$, and type, $F(3,117) = 91.18$, $MSe = 2.88$.

In comparison to the non-review control condition that utilized distributed test-study repetitions, significantly higher recall was found for reviews spaced across days (expanding and uniform conditions, $F(1,27) = 49.16$, $MSe = 3.85$) and for reviews given immediately after initial learning (massed condition, $F(1,13) = 5.51$, $MSe = 4.73$). Spacing effects were also shown for the test-study and study-only condition, as delayed reviews (expanding and uniform conditions) were found to be significantly more effective than immediate reviews (massed condition) (test-study: $F(1,39) = 14.69$, $MSe = 5.60$; study-only: $F(1,39) = 23.23$, $MSe = 5.91$). For the test-only condition, no significant difference was found between delayed reviews and immediate reviews. Consistent with the previous two experiments, no advantage was found for the expanding spacing of reviews over the uniform distributed spacing. In fact, the uniform condition generally performed better than the expanding condition in Experiment 3, and this difference approached significance, $F(1,26) = 3.85$, $MSe = 9.22$, $p = 0.061$.

The third panel in Figure 1 shows that testing once again benefited learning in addition to the effect of distributed practice. For both the expanding and uniform spacings of delayed reviews, the test-study condition produced significantly higher recall than the study-only condition, $F(1,27) = 4.44$, $MSe = 1.30$. The effect sizes for testing were small to moderate for the expanding ($d = 0.28$) and uniform spacings ($d = 0.33$). These effects were possibly weakened somewhat by a performance ceiling in the test-study condition. For immediate reviews (massed condition), testing effects were found for both the test-study, $F(1,13) = 6.47$, $MSe = 2.07$, and test-only conditions, $F(1,13) = 5.90$, $MSe = 3.20$. The effect size for testing was moderate to large ($d = 0.70$).

It is interesting that the massed review condition was very similar to the immediate review condition used in Experiment 2 except for the longer retention interval; yet study-only review was more effective than test-only review in Experiment 2 and test-only review was more effective than study-only review for the massed review condition in Experiment 3. It is also interesting that the test-only condition was just as effective

as the test-study condition for massed review. These results suggest that the benefits of testing appear to be longer lasting than the benefits of study.

EXPERIMENT 4

To examine the possible influence that longer retention intervals may have on the effects of distributed practice, testing, and expanding test spacing, a longer retention interval was used in Experiment 4. The retention interval between the last review session and the final test was increased from 3 days in Experiment 3 to 8 days in Experiment 4; otherwise, Experiments 3 and 4 were very similar. The increased retention interval provided a look at these effects of reviews across a time period that was more consistent with educators' goals (Bahrick *et al*., 1993; Bahrick and Hall, 1991). It was expected that the benefits of distributed practice and testing shown in Experiment 3 would be maintained across this long retention interval.

Method

Participants

A total of 42 students who were enrolled in an upper-division psychology course at Loyola University of Chicago participated in the experiment as part of a course requirement. No participant had participated in any of the earlier experiments.

Materials and procedure

The materials and procedures for Experiment 4 were very similar to those of Experiment 3, but some differences existed. First, the spacings between review sessions were identical to those used in Experiment 3 (1 day, 2 days, and 3 days for expanding, 2 days, 2 days, and 2 days for uniform, and 0 days, 0 days, and 0 days for massed), but the delay between the final review session and the final test was increased to 8 days for the expanding and uniform conditions and 14 days for the massed condition. This was done to investigate the benefits of review tests across a longer interval than that used in Experiment 3. Second, the procedure for review sessions was modified. Rather than providing booklets that randomly presented the items, these items were presented using two full sheets of paper: a test-sheet and a study-sheet. On the test-sheet, the cue members of the test-only and test-study items were presented in a mixed order with a blank line provided after each cue. Participants were instructed to provide a definition for as many of the words as they could. A study-sheet was stapled close to the back of the test-sheet. Participants were explicitly told not to write on the test-sheet once the study sheet had been opened. On the study-sheet, the cue words with their definitions were presented in a mixed order for the study-only and test-study conditions. The same words were presented as study-only, test-only, and test-study items on each of the three reviews, but the order of presentation changed. As in Experiment 3, 10 items were randomly assigned to each of the four review types (test-study, study-only, test-only, and nonreviewed) and all 40 words that had been originally learned (30 reviewed, 10 not reviewed) were presented on the final paper-and-pencil, cued-recall test. The same random order of items was used for all participants, and participants were given as long as they needed to complete the test.

Results and discussion

Results for the final cued-recall test of Experiment 4 are summarized at the bottom of Table 1. A 4 × 3 mixed ANOVA was used to analyse these results. Results showed significant main effects of type of review, $F(3,117) = 60.96$, $MSe = 3.07$, and spacing, $F(2,39) = 9.72$, $MSe = 15.02$, and a significant interaction between these variables, $F(6,117) = 6.85$, $MSe = 3.07$.

As was the case in Experiment 3, delayed reviews (expanding and uniform conditions) produced significantly greater learning than the non-review control group (expanding: $F(1,13) = 8.26$, $MSe = 8.37$; uniform: $F(1,13) = 39.62$, $MSe = 3.35$). Immediate reviews (massed condition) also produced significantly higher recall than the nonreview control group for the test-study and test-only conditions, $F(1,13) = 6.45$, $MSe = 2.44$, but not for the study-only condition. Spacing effects were found for all three review types. The expanding and uniform distributed spacing produced significantly higher recall than the massed spacing in the test-study, $F(1,39) = 27.22$, $MSe = 4.11$, and study-only conditions, $F(1,39) = 16.20$, $MSe = 6.67$, and only the uniform spacing was significantly higher than the massed spacing in the test-only condition, $F(1,39) = 9.57$, $MSe = 8.60$. Once again, no significant differences were found between the expanding spacing and the uniform spacing across review types.

The fourth panel of Figure 1 shows that in Experiment 4, testing had its largest benefit for learning beyond that provided by distributed practice. The test-study condition produced significantly higher recall than the study-only condition did for all three spacing schedules (expanding: $F(1,13) = 8.39$, $MSe = 4.09$; uniform: $F(1,13) = 9.03$, $MSe = 2.88$; massed: $F(1,13) = 53.18$, $MSe = 0.60$). The size of these effects were very large, as Cohen's d ranged from 1.09 to 1.30. For the massed spacing, the test-only condition also produced significantly higher recall than did the study-only condition, $F(1,13) = 5.51$, $MSe = 2.59$. These results suggest that delayed review with testing is especially important for longer-term retention of information.

GENERAL DISCUSSION

Benefits of testing beyond that provided by distributed study were found in all four experiments. Moderate to large benefits of testing were found for test repetitions in Experiment 1, for immediate reviews in Experiment 2, and for delayed reviews in Experiment 4. A smaller effect of testing was shown in Experiment 3, which may be related to a possible ceiling effect or to the shorter 3-day retention interval that was used. The practical benefits of these effects are highlighted by the fact that testing benefits were obtained in addition to the already sizable benefits of distributed study. Results of the current study support the conclusion that distributed tests are a valuable tool for educators that should be utilized regularly in the classroom for their direct impact on learning (Dempster, 1996).

The current study results especially highlight the beneficial effects of reviews occurring in separate learning sessions that are spaced days apart from each other. In Experiments 3 and 4, three different timings of reviews were directly compared with each other; the nonreview control conditions utilized test-study repetitions within a single learning session (similar to Experiment 1's method), the massed conditions

provided test-study repetitions plus review immediately following initial study (similar to Experiment 2's method), and the expanding and uniform spaced conditions provided test-study repetitions plus review spaced across days. The large advantages found for the spaced conditions in comparison to the control and massed conditions suggest that although test-study repetitions within initial learning or test-study reviews occurring immediately after initial learning can be effective, review tests spaced days apart from each other are the most effective.

Explanation of testing effects

Bahrick and Hall (1991) suggested that tests promote learning by providing either preventive or corrective maintenance of information. Preventive maintenance of information results when information is successfully retrieved at the time of testing, and this produces more durable future access to the remembered information. Corrective maintenance, on the other hand, occurs when information is not accessible at the time of the review test, but the process of attempting retrieval or reviewing the information restores access to the lost information. Examining the differences between test-only, study-only, and test-study review in their abilities to provide preventative or corrective maintenance provides a theoretical framework to understand better how retrieval benefits learning.

Test-only review presents the learner with a challenge by providing only partial information to the learner at the time of the test. If the learner is successful in retrieving the missing information, the effort put forth to locate the information in memory is rewarded by decreased forgetting or increased preventive maintenance of that information. However, test-only review's potential for corrective maintenance is limited because the learner isn't given correct-answer feedback to help restore information in memory following unsuccessful recall efforts. Research on hypermnesia or reminiscence has demonstrated, however, that even in the absence of test feedback, corrective maintenance can arise from the process of searching memory or from spontaneous recovery of information (Payne, 1987; Wheeler, 1995). In the current studies, an examination of hypermnesia for the test-only conditions did not show significant improvements in overall recall from the time of the first review to the time of the final test in any of the experiments. Thus, the benefits of test-only review appear to have been primarily from reducing the forgetting of remembered information rather than from regaining access to forgotten information.

The study-only condition continually re-presents complete item information to the learner during review and thus, increases dramatically the potential for corrective maintenance. This corrective maintenance advantage for the study-only condition, however, did not always lead to superior performance in comparison to test-only review. In Experiment 1, Experiment 2, and the spaced conditions of experiment 4, for example, the corrective maintenance advantage of the study-only condition must have been offset by a preventive maintenance advantage favoring the test-only condition, since no significant differences were apparent between the study-only and test-only conditions. Moreover, the test-only condition produced significantly higher overall recall levels than did the study-only condition in the massed conditions of Experiment 3 and 4. Apparently, the preventive maintenance advantage for the test-only condition exceeded the corrective maintenance advantage for the study-only condition in these situations. These situations were characterized by a fairly short

interval between initial study and review that made successful recall at review probable and a long interval between review and final recall that made durable access to information necessary.

There are several possible explanations of how testing produces a greater amount of preventive maintenance than a comparable amount of additional study. Some explanations have focused on encoding and have suggested that the difficulty imposed by testing compels the learner to process information more thoroughly than study (Cuddy and Jacoby, 1982; Jacoby, 1978). This leads to a stronger memory trace or to greater integration of that information with previous knowledge. Others have focused on the storage of information and have suggested that the act of successful retrieval modifies the memory trace and helps to consolidate the information in memory (Bjork, 1975). Yet other explanations have suggested that testing provides the learner with practice at retrieving the information which facilitates later retrieval efforts in similar contexts (Carrier and Pashler, 1992). A variation of this explanation is that retrieval practice provides the learner with feedback about the quality of existing connections in memory and that new modifiers are created when necessary (King *et al.*, 1980). Differences in theoretical opinions about encoding and retrieval effects are notoriously difficult to reconcile (Watkins, 1990), and it is very possible that testing benefits both.

Test-study retains the preventive maintenance advantage of the test-only condition while also providing the corrective maintenance potential of the study-only condition. As shown across all four experiments, combining these benefits together results in a highly effective learning situation. Part of the effectiveness of test-study review also may be related to a possible carry-over or potentiation effect arising from the connected presentation of the test and study phases (Izawa, 1970). For example, a learner who is unsuccessful in remembering the tested information in the test phase may enter the study phase with increased motivation to learn, with memory priming for the to-be-relearned information, or with an altered strategy for how best to encode the information. This is consistent with previous studies of metacognition that have shown that testing improves learners' assessments of their learning and their ability to allocate study time appropriately (Cull and Zechmeister, 1994; King *et al.*, 1980; Spellman and Bjork, 1992). Thus, the amount of corrective maintenance provided by the study phase of test-study review may exceed that of study-only review.

Similarly, the amount of preventive maintenance provided by the test phase of test-study review may exceed that of test-only review. In the test-study condition, the study phase provides the learner with feedback concerning whether a test response was successful which is absent for the test-only condition. This provides the learner with confirmation that an answer is right if the learner indeed answered correctly. This confirmation may help consolidate information better in memory and thereby increase preventive maintenance. Regardless of whether the benefits of test and study are accentuated or merely combined in test-study review, the success of this restudy method is large and should be utilized by educators.

The effectiveness of test-study review shown in the current experiments is also consistent with Glover's (1989) observation that review tests can vary in the completeness of information processing that they induce and that different test types may produce different levels of later recall. Glover showed, for example, that recall tests are more effective in producing later learning than are recognition tests regardless of whether the final test is recall or recognition. In the current experiments, test-study

review can also be interpreted as eliciting more complete information processing than does study-only or test-only review. This is especially interesting in comparison to study-only review because the study-only condition allows more time with total item information available, yet the test-study condition was consistently more effective. The success of the test-study condition shows that having less information can be better for the learner in certain situations, depending on the opportunities provided for and the demands placed upon the learner (Schmidt and Bjork, 1992).

The spacing of testing was also shown to play an important role for testing effectiveness. This is consistent with previous studies examining test timing (Glover, 1989; Landauer and Bjork, 1978). In Experiments 1, 3, and 4, tests that are massed together were not as effective as spaced tests. According to a completeness of processing interpretation of these results, this occurred because the shorter gap between initial study and restudy in the massed conditions relative to the spaced conditions reduced the challenge that was placed upon the learner who was trying to reconcile the targeted information with previous knowledge. Thus, less complete information processing occurred. The reduction in processing was most obvious in the massed condition of Experiment 1, which was the only learning situation where the test-study condition was not significantly more effective than the study-only condition. In this situation, information was re-presented to the learner without any time intervening between study and restudy, making it highly likely that the information was continuously available in short-term memory (see Peterson and Peterson, 1959; Zechmeister and Nyberg, 1982). For testing to be most effective, the test must force the learner to retrieve information from long-term rather than short-term memory.

Finally, the benefits of testing were amplified as the delay between the last review and the final test increased. A comparison of the uniform-spaced test-study and study-only conditions shows, for example, that as the retention delay increased from 3 days in Experiment 3 to 8 days in Experiment 4, the advantage for the test-study condition increased from 7 percentage points (98% minus 91%) to 25 percentage points (89% minus 64%). The greater forgetting with longer retention intervals for the study-only condition was also apparent in comparison to the test-only condition. With the 3-day retention interval (Experiment 3), the study-only condition was more effective than the test-only condition by a sizeable 28 percentage points (91% versus 64%), and with the 8-day retention interval (Experiment 4) the pattern was reversed as the study-only condition was less effective then the test-only condition by 5 percentage points (64% versus 69%). Although a ceiling effect possibly can explain the larger testing effect with the longer retention interval for the test-study condition, it cannot explain the similar effect for the test-only condition. Additional tests of the tendency for test effects to increase as the retention interval lengthens needs to be made within the same experiment, since longer retention intervals are very consistent with the goals of education.

Reconsidering the expanding-test-series mnemonic

Expanding the test spacing, however, did not independently benefit learning in any of the learning situations studied. Across the four experiments, expanding test spacing did not provide significantly higher levels of learning than those provided by uniform distributed test spacing. This finding is consistent with findings by Cull *et al*. (1996) in showing no advantage of expanding distributed spacing over uniform distributed

spacing for tests with feedback, but is inconsistent in showing no advantage for tests without feedback.

The theoretical goal of expanding testing is to maximize the preventive maintenance of information by maintaining high rates of successful retrieval throughout the test pattern and gradually increasing the difficulty of retrieving item information in order to strengthen storage. Optimal preventive maintenance is thought to occur when information is successfully retrieved just as it is on the verge of being forgotten (Baddeley, 1990; Landauer and Bjork, 1978). For the study-only and test-study conditions, it is unclear whether letting information slip past the forgetting point is necessarily worse for long-term information retention because corrective maintenance of information is still made available through information re-presentation. The equivalence of the expanding and uniform conditions when item feedback is available following testing may reflect a boundary of expanding testing's effectiveness.

Because of the test-only condition's greater reliance on preventive maintenance than the test-study or study-only conditions, expanding test benefits definitely were expected for those conditions. The failure to find expanding test-series effects for the test-only conditions were surprising and may be related to the specific expanding intervals chosen in the current studies. The intervals were chosen so that the first test would be easier and the last test would be harder in the expanding rather than in the uniform condition, while the total amount of delay between tests would remain the same. The first review tests in the expanding condition may not have been close enough, though, to provide a qualitatively easier first test than the uniform condition. Perhaps, a pattern of review tests where the first test is given minutes or hours from initial study with later tests spaced days apart may be more effective for test-only review than the 1 day, 2 days, 3 days pattern used in the current studies. Attempts to tailor expanding patterns to each individual learner's own forgetting rates may also prove to be effective.

Conclusion

Distributed study and testing were shown to have reliable and robust effects on learning that can be combined to provide an extremely powerful learning aid. Educators should strive to incorporate distributed testing regularly in their teaching. For instance, in lectures, teachers can incorporate testing by inserting questions about topics that were covered earlier that day or on previous days. Teachers may already do this as a means to gain students' attention or orient learners to a topic area, but teachers should make a point to allow learners enough time to search for the answer so that the test can also effectively promote learning. Also, providing review quizzes targeting important material should improve students' understanding of that information in addition to providing a method to evaluate performance. Future research needs to be conducted to ensure that distributed testing is effective in real-word education settings and that teachers are utilizing this learning tool.

AUTHOR NOTE

These experiments were conducted as part of a dissertation submitted by the author to the graduate school, Loyola University of Chicago. The author thanks Eugene

Zechmeister, dissertation committee chairman, and Tricia Tenpenny, Lois Leidahl, and Emil Posovac for many comments and suggestions that helped shape the preparation of the final draft of this article. Some results from these experiments were presented at the 1995 meeting of the Midwestern Psychological Association.

REFERENCES

Allen, G. A., Mahler, W. A. and Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, **17**, 573–585.

Baddeley, A. (1990). *Human memory: theory and practice*. Boston, MA.: Allyn & Bacon.

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

Bahrick, H. P. and Hall, L. K. (1991). Preventive and corrective maintenance of access to knowledge. *Applied Cognitive Psychology*, **5**, 1–18.

Bartlett, J. C. (1977). Effects of immediate testing on delayed retrieval: Search and recovery operations with four types of cue. *Journal of Experimental Psychology: Human Learning & Memory*, **3**, 719–732.

Bartlett, J. C. and Tulving, E. (1974). Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval. *Journal of Verbal Learning and Verbal Behaviour*, **13**, 297–309.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), In *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ.: Erlbaum.

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

Carrier, M. and Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, **20**, 633–642.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155–159.

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

Cull, W. L. and Zechmeister, E. B. (1994). The learning ability paradox: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, **22**, 249–257.

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

D'Anna, C. A., Zechmeister, E. B. and Hall, J. J. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior*, **23**, 109–122.

Darley, C. F. and Murdock, B. B. J. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, **91**, 66–73.

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. (1992). Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education*, **25**, 213–217.

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

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

Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, **10**, 577–585.

Glover, J. A. (1989). The 'testing' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, **81**, 392–399.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories of cognitive psychology: The Loyola symposium*. Hillsdale, NJ.: Erlbaum.

Hogan, R. M. and Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, **10**, 562–567.

Izawa, C. (1970). Optimal potentiating effects and forgetting: Prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, **83**, 340–344.

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

King, J. F., Zechmeister, E. B. and Shaughnessy, J. J. (1980). Judgments of learning: The influence of retrieval practice. *American Journal of Psychology*, **93**, 329–343.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review*, **76**, 82–96.

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

Landauer, T. K. and Eldridge, L. (1967). Effects of tests without feedback and presentation-test interval in paired-associate learning. *Journal of Experimental Psychology*, **75**, 290–298.

LaPorte, R. E. and Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, **67**, 259–266.

Madigan, S. A. and McCabe, L. (1971). Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **10**, 101–106.

McDaniel, M. A., Kowitz, M. D. and Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition*, **17**, 423–434.

McDaniel, M. A. and Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 371–385.

McDaniel, M. A. and Pressley, M. (1984). Putting the keyword method in context. *Journal of Educational Psychology*, **76**, 598–609.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, **9**, 596–606.

Modigliani, V. (1976). Effects on later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory*, **2**, 609–622.

Nelson, T. O. and Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 676–686.

Nungester, R. J. and Duchastel, P. C. (1982). Testing versus review: Effects on Retention. *Journal of Educational Psychology*, **74**, 18–22.

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin*, **101**, 5–27.

Peterson, L. R. and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**, 193–198.

Rea, C. P. and Modigliani, V. (1985). the effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning*, **4**, 11–18.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, **11**, 641–650.

Runquist, W. N. (1986a). Retrieval specificity and the attenuation of forgetting by testing. *Canadian Journal of Psychology*, **40**, 84–90.

Runquist, W. N. (1986b). Changes in the rate of forgetting produced by recall tests. *Canadian Journal of Psychology*, **40**, 282–289.

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

Sones, A. M. and Stroud, J. B. (1940). Review with special reference to temporal position. *Journal of Educational Psychology*, **31**, 665–676.

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

Underwood, B. J. (1970). A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, **9**, 573–580.

Underwood, B. J. (1982). Paired associate learning: Data on pair difficulty and variables that influence difficulty. *Memory & Cognition*, **10**, 610–617.
Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist*, **45**, 328–335.
Wenger, S. K., Thompson, C. P. and Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 135–144.
Wheeler, M. A. (1995). Improvement in recall over time without repeated testing: Spontaneous recovery revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 173–184.
Whitten, W. B. and Bjork, R. A. (1977). Learning from tests: effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, **16**, 465–478.
Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology*, **8**, 58–81.
Zechmeister, E. B. and Nyberg, S. E. (1982). *Human memory: An introduction to research and theory*. Monterey, CA.: Brooks/Cole.

