Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 57 (2007) 151–162
Journal of Memory and Language
www.elsevier.com/locate/jml

Repeated retrieval during learning is the key to long-term retention ☆

Jeffrey D. Karpicke *, Henry L. Roediger III

Department of Psychology, Washington University, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899, USA

Received 11 July 2006; revision received 9 September 2006
Available online 13 November 2006

Abstract

Tests not only measure the contents of memory, they can also enhance learning and long-term retention. We report two experiments inspired by Tulving's (1967) pioneering work on the effects of testing on multitrial free recall. Subjects learned lists of words across multiple study and test trials and took a final recall test 1 week after learning. In Experiment 1, repeated testing during learning enhanced retention relative to repeated studying, although alternating study and test trials produced the best retention. In Experiment 2, recalled items were dropped from further studying or further testing to investigate how different types of practice affect retention. Repeated study of previously recalled items did not benefit retention relative to dropping those items from further study. However, repeated recall of previously recalled items enhanced retention by more than 100% relative to dropping those items from further testing. Repeated retrieval of information is the key to long-term retention.

Keywords: Testing effect; Retrieval processes; Learning; Retention

Students often learn new material by repeatedly studying and testing themselves over the material and monitoring their progress in learning. This kind of multiple-trial learning was formerly a principal method used by experimental psychologists to study human learning and memory in many different paradigms. In typical multitrial free recall experiments, subjects study a list of items during a study trial, recall as many of the items as they can during a test trial, restudy the list in another study trial, recall in another test trial, and so on, across alternating study and test trials (i.e., study–test–study–test, or STST). Learning is measured as the increase in retention across trials. Ever since Ebbinghaus (1885/1964) observed that recall increased across repeated study and test trials, research has been devoted to understanding the effects of repetition on learning (for a review, see Greene, 1992, Ch. 7), and by the 1960s the effects of several variables on multitrial learning had been investigated (see Tulving, 1968). However, with a few exceptions (e.g., Dunlosky & Hertzog, 1997; Koriat, Sheffer, & Ma'ayan, 2002), multitrial learning is rarely investigated in current memory research, even though it may reflect how students master new information in school, such as multiplication facts or foreign language vocabulary.

☆ We thank Julie Evans, Liz Rhoades, Barbara Russo, and Patrick Weaver for assistance with data collection and scoring. We also thank Daniel Burns, Michael Masson, James Nairne, and Endel Tulving for comments on the manuscript. This research was supported by grants from the James S. McDonnell Foundation and the Institute of Education Sciences.
* Corresponding author. Fax: +1 314 935 7588.
E-mail address: karpicke@wustl.edu (J.D. Karpicke).

0749-596X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2006.09.004

Debate concerning the nature of multitrial learning swirled in the 1950s and 1960s, with some theorists arguing that learning occurred as an incremental process and others arguing that learning was an all-or-none process (see Crowder, 1976, Ch. 8). Learning appears to occur gradually across study and test trials, as evidenced by the negatively accelerated learning curves found in virtually all experiments. To explain the growth of the learning curve, the incremental position held that learning reflected the increase in the strength of individual memory traces. Each time the to-be-learned items were presented for study, the memory traces for those items gained some quantity of strength, and once they reached a certain threshold of strength they were recalled on the test trial. The incremental position, in place at least since Ebbinghaus, was challenged in the 1950s and 1960s by researchers arguing that learning was instead an all-or-none process (e.g., Rock, 1957). The all-or-none position held that each item remained in one of two discrete states, learned or not learned, and that each presentation of an item either caused it to transition into the learned state or failed to do so. Thus the apparent gradual nature of learning seen in learning curves was simply an artifact of averaging across learned and not-learned items in a list.

We bother to remind readers of this debate to point out an important (but implicit) assumption about learning that is shared in both the incremental and all-or-none positions, as well as virtually all theories of learning. The common assumption is that learning or acquisition of the list happens during study trials and that test trials serve only as an opportunity for subjects to express what they have learned. Tests are neutral events that assess learning but do not affect it. Of course, these assumptions about studying and testing are made not just in memory experiments; we would hazard the guess that most teachers and professors, as well as most students, think of learning as a process that occurs during study of material (readings, lectures, and study groups) and that testing serves merely the purpose of assessing what was learned. Again, tests are assumed to be relatively neutral events in the learning process.

According to these assumptions, we can make predictions about what should happen with other types of study and test sequences besides the common STST multitrial paradigm. For purposes of exposition (and following Tulving, 1967), we will refer to a sequence of four sequential study or test events as a cycle, so STST would constitute a cycle of alternating study and test events. Relative to this baseline, increasing the number of study trials during a learning cycle (e.g., having students study material three times and take one test, or SSST) should enhance learning relative to the standard condition of alternating study and test trials (STST) when subjects in both conditions are tested on the fourth (test) event common to both cycles. This outcome is predicted because students study the material an additional time in the SSST case. On the other hand, increasing the number of test trials in a cycle at the expense of study trials (e.g., STTT) should have a negative effect on learning by reducing the number of study trials relative to the standard condition.

Forty years ago, Tulving (1967) investigated precisely this issue and reported the results in this journal. He had subjects learn a list of 36 words under standard (STST), repeated study (SSST), or repeated test (STTT) conditions. Each study or test trial lasted 36 s so that the total time spent was held constant across conditions, and oral recall was used during the test phase. The conditions involved 6 cycles, so altogether subjects in the repeated study condition studied the list 18 times and took six recall tests, whereas subjects in the standard condition studied the list 12 times and took 12 tests, and subjects in the repeated test condition studied the list only six times and took 18 recall tests. If subjects simply acquire the list items during study trials and then express their knowledge during test trials, then increasing the number of study trials from 6 to 12 to 18 should produce large positive effects on learning.

On the contrary, Tulving (1967) observed generally equivalent learning under the three different conditions. Increasing the number of study trials in the SSST condition did not boost learning, and although the repeated test (STTT) group recalled slightly fewer items overall than subjects in the other two conditions, the difference between the conditions decreased across cycles. Tulving's results clearly showed that tests are not simply a neutral assessment of what has been learned but also produce learning, perhaps as much learning as during a study trial. His experiment showed what is now called the testing effect, the finding that tests not only assess learning but also enhance it (a fact uncovered early in the 20th century using different methods, e.g., Gates, 1917, among others). Subsequent studies of the testing effect also showed that testing often improves long-term retention relative to additional studying (see Hogan & Kintsch, 1971; McDaniel & Masson, 1985; Roediger & Karpicke, 2006b). For example, Hogan and Kintsch (1971) had subjects study a list of words four times (SSSS) or study it once and recall it three times (STTT). The conditions were similar to one cycle in Tulving's (1967) experiment. On a final free recall test 2 days later, the repeated test group recalled more than the repeated study group. Even though the repeated study group restudied the entire list three times while the test group could only re-experience whatever they could recall on the three tests, testing led to better long-term retention than studying (see also Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonanno, 2003). Roediger and Karpicke (2006a) provide a recent review of the literature on the testing effect.

Tulving's (1967) experiment sparked a burst of research in the late 1960s and early 1970s (e.g., Bregman & Wiener, 1970; Donaldson, 1971; Lachman & Laughery, 1968; Patterson, 1972; Rosner, 1970) that replicated his basic findings, but has received little attention since. In the present experiments, we investigated the effects of repeated studying and testing during multitrial learning and on long-term retention (measured after a week delay) using conditions similar to Tulving's. Tulving's results suggested that a test trial was as good as a study trial for improving learning (see also Lachman & Laughery, 1968), but based on other findings in the testing effect literature we predicted that test trials would actually enhance long-term retention more than study trials. We examined recall after a long delay (1 week) in order to measure the enduring effects of the three types of cycles (for reasons that will be described in a few paragraphs).

Tulving (1967) and others of that era used recall tests that immediately followed study trials, which mixes together contributions from primary and secondary memory on the test (Glanzer & Cunitz, 1966). When SSST, STST, and STTT cycles are all compared on the fourth trial common to all groups, the STTT condition may be at a disadvantage relative to the other two; subjects with an immediately preceding study trial may retrieve from primary memory whereas those given repeated tests cannot (on the last, common test). In our experiment, we sought to separate contributions from primary and secondary memory during the tests by using a correction procedure developed by Tulving and Colotla (1970) that classifies each recalled word as retrieved from primary or secondary memory (see also Watkins, 1974).

In Experiment 1, we compared SSST, STST, and STTT cycles and had subjects participate in 5 cycles (i.e., 20 study or test events). The main purpose of Experiment 1 relative to Tulving's (1967) original work was to provide a 1-week delayed test. Although Tulving's (1967) research suggested that as much learning accrues from a test trial as from a study trial, more recent work suggests that test trials may actually create greater learning as assessed on long term tests than do study trials. One possible theoretical reason for this prediction comes from the principle of transfer appropriate processing: If the criterial test involves retrieving information with minimal cues after a delay, then prior practice at such retrieval should provide more benefit than repeated study (e.g., Roediger, Gallo, & Geraci, 2002). In addition, several experiments produced results suggesting that repeated testing produces greater gains than repeated studying (e.g., Roediger & Karpicke, 2006b; Thompson et al., 1978; Wheeler et al., 2003).

In a second experiment, we took a different approach to examining the effects of repeated studying vs. testing on the initial rate of learning and later retention. In one condition in Experiment 2, when an item was recalled on a test, it was dropped from further study trials; however, subjects were still required to recall the entire list of items on the test (see Thompson et al., 1978). In another condition, recalled items were dropped from further study trials and also from further test trials by instructing subjects to recall only the words studied on the previous study trial. Thus, all words were recalled one time before they were dropped in this condition. Compared to the standard study-test condition, in which all items are studied and tested on all study and test trials, these new conditions allowed us to investigate more cleanly the effects of repeated studying and repeated testing on retention.

Experiment 1

In Experiment 1, subjects learned a list of words under standard (STST), repeated study (SSST), or repeated test conditions (STTT). The subjects were then given a final free recall test 1 week after the learning phase. We predicted that even though initial learning might be similar or equivalent under the three conditions (as Tulving, 1967, observed), repeated testing would lead to superior retention on the delayed test given a week later. That is, a test trial may produce as much initial learning as a study trial, but test trials should confer greater benefit to long-term retention than study trials. This pattern of results would be consistent with other literature on the testing effect showing that additional studying sometimes benefits learning in the short-term, but that testing during learning leads to better retention after a delay (see Roediger & Karpicke, 2006a, 2006b). For example, Roediger and Karpicke (2006b) found that repeated study led to an advantage over repeated testing on an immediate test, but that repeated testing produced superior long-term retention after 2-day and 1-week delays.

Methods

Subjects

Sixty Washington University undergraduates, ages 18–26, participated in exchange for course credit.

Materials

Forty unrelated words were selected from the norms of Paivio, Yuille, and Madigan (1968). The words ranged in frequency from 29 to 50 words per million.

Design

Twenty subjects were assigned to each of the three learning conditions: Standard (STST), repeated study (SSST), or repeated test (STTT). In the standard condition, subjects studied and recalled the list of words during alternating study and test trials. In the repeated

study condition, subjects studied the list three times and took one recall test in each cycle of four trials. In the repeated test condition, subjects studied the list once and took three consecutive recall tests in each cycle. The learning phase involved 5 cycles, so subjects studied and recalled the words during a total of 20 trials, with the number of study and test trials being 10/10 in the standard condition, 15/5 in the repeated study condition, and 5/15 in the repeated test condition.

Procedure

The subjects were tested in groups of five or fewer. They were told that they would study and recall a list of words during several trials in the learning phase. At the beginning of each study trial, a "Ready" prompt was shown on the computer screen for 1 s, and then the 40 words were presented on the screen at a rate of 3 s per word (the words were presented in a different random order on each study trial). The subjects were told to study the words so that they could recall them on the test trials. The beginning of each test trial was indicated by a tone (presented over headphones) and a "Recall" prompt that remained on the computer screen throughout the test. During each test trial the subjects were given 2 min to write down as many of the words as possible, in any order, in a response booklet. Although Tulving (1967) and others who followed him used shorter amounts of recall time (e.g., 36 s to recall 36 words in Tulving, 1967) we gave subjects 2 min tests to ensure that they had enough time to express everything they could recall (see Roediger & Thorpe, 1978). The transition from one test trial to another (in the repeated test condition) was indicated by a tone as well as a change in the background color on the computer screen: The background was blue during the first test, green during the second test, and red during the third test. At the end of each test trial, the subjects were instructed to turn to the next page in their response booklets and they were told not to look back at any of their previous responses at any time during the learning phase.

The subjects returned for the final test 1 week after the learning phase. They were given 10 min to write down as many of the words as they could recall, in any order, and were also instructed to draw a line on their recall sheet to mark their progress at one minute intervals (Roediger & Thorpe, 1978). This procedure allowed us to measure cumulative recall and to ensure that subjects had exhausted their knowledge by the end of the 10 min recall test. At the end of the final test the subjects were debriefed and thanked for their participation.

Results

Primary memory

Recall from primary memory did not differ in the three conditions ($F < 1$), averaging 1.54 items collapsed across conditions. Primary memory recall also remained constant across cycles in the learning phase ($F < 1$). The contribution of primary memory to recall was low relative to other studies using this estimation procedure (see Watkins, 1974). However, in multitrial free recall, subjects often recall newly learned items prior to the items that they have recalled correctly on previous trials (Battig, Allen, & Jensen, 1965), and this effect may have attenuated the contribution of primary memory in this experiment relative to standard studies of single-trial free recall where estimates of primary memory are about 3–3.5 items (Watkins, 1974). The remaining analyses were carried out only on recall from secondary memory, with primary memory recall removed from the data, although we should note that the same conclusions were obtained when we analyzed the data without removing primary memory recall.

Learning phase

Fig. 1 shows the mean proportion of words recalled on each trial from secondary memory. The figure shows that recall increased in a regular fashion in all three conditions in a fairly similar manner, although the standard condition produced the best learning over trials. In addition, the repeated study and repeated test conditions interacted across cycles: Repeated studying led to better recall in the first 2 cycles (trials 4 and 8) but was eclipsed by the repeated testing condition in the third cycle and then surpassed in later cycles (trials 16 and 20). A 3 (learning condition: STST, SSST, or STTT) × 5 (cycle: 1–5) ANOVA performed on the mean proportion of words recalled during the fourth trial in each cycle revealed a main effect of cycle, $F(4, 228) = 450.18$, $\eta_p^2 = .89$ (which simply indicates that learning occurred across cycles), as well as a main effect of learning

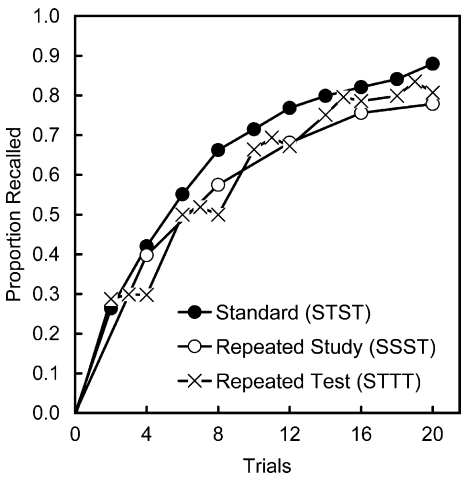

Fig. 1. Proportion of words recalled during the learning phase in Experiment 1.

condition, $F(2, 57) = 5.40$, $\eta_p^2 = .16$, and a learning condition × cycle interaction, $F(8, 57) = 5.12$, $\eta_p^2 = .15$. (All effects deemed significant meet the $p < .05$ level of significance.) Early in learning, on trial 4, the repeated study condition outperformed the repeated test condition (40% vs. 30%, $d = 1.01$, $p_{rep} = .98$), as did the standard condition (42% vs. 30%, $d = 1.56$, $p_{rep} = .99$), though the standard and repeated study conditions were only slightly different (42% vs. 40%, $d = .20$, $p_{rep} = .67$). (p_{rep} is an estimate of the probability of replicating the direction of an effect, described by Killeen, 2005.) However, by the end of the learning phase (on trial 20) recall in the repeated test condition was slightly better than recall in the repeated study condition (81% vs. 78%, $d = .23$, $p_{rep} = .70$), and the standard condition outperformed both the repeated test (88% vs. 81%, $d = .81$, $p_{rep} = .96$) and repeated study conditions (88% vs. 78%, $d = .93$, $p_{rep} = .98$).

The potentiating effect of testing

Experiment 1 also revealed another interesting consequence of testing, called the potentiating effect of testing by Izawa (1971), which is that a test trial increases learning on the next study trial. Repeated testing in the STTT condition increased the number of items acquired on a subsequent study trial relative to the number of items acquired after taking a single test in the STST condition. The potentiating effect of repeated testing is shown in Fig. 2, which shows performance on the tests that followed each of the first 5 study trials in the STST and STTT conditions. For the STTT condition, Fig. 2 shows performance averaged across the three test trials within each cycle for all five cycles (because there were only five study trials). For the STST condition, comparable data are presented from the first 5 test trials (or the first half of the entire set of 10 study and test trials). This comparison reveals that the number of new items recalled following the second, third, fourth, and fifth study trials increased when subjects had taken 3 tests vs. 1 test before the study trial. Recall of new items was the same in both conditions on the first opportunity to measure recall (following the first study trial). However, recall was greater in the STTT condition following the second study trial, and the effect persisted for study trials 3–5. Thus, repeated testing potentiated the acquisition of items on a subsequent study trial relative to taking a single test.

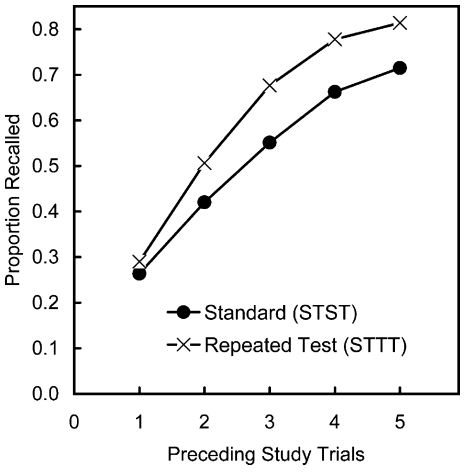

Fig. 2. The potentiating effect of testing. The proportion of items recalled on tests that followed the first 5 study trials in the STTT and STST conditions. Recall in the STTT condition is the average of performance on the three tests within each cycle.

Final recall

Fig. 3 shows cumulative recall on the 10-min final free recall test given 1 week after the learning phase. The figure shows that from the very first minute of the recall period the repeated study condition performed worse than the other two conditions. By the end of the 10 min recall period, subjects recalled 68% of the words in the standard condition and 64% in the repeated test condition, though this was only a small effect ($d = .23$, $p_{rep} = .70$), but both conditions led to better final recall than the repeated study condition. Subjects in the repeated test condition outperformed those in the repeated study condition (64% vs. 57%, $d = .44$, $p_{rep} = .83$) and subjects in the standard condition recalled more than those in the study condition (68% vs. 57%, $d = .70$, $p_{rep} = .93$). The result suggests that conditions with more frequent testing led to better long-term retention.

Fig. 4 shows a conditional analysis of the probability of final recall as a function of the number of times an item was recalled during the initial learning phase. This conditional analysis is correlational in nature, of course, and subject to item-selection effects, but nonetheless

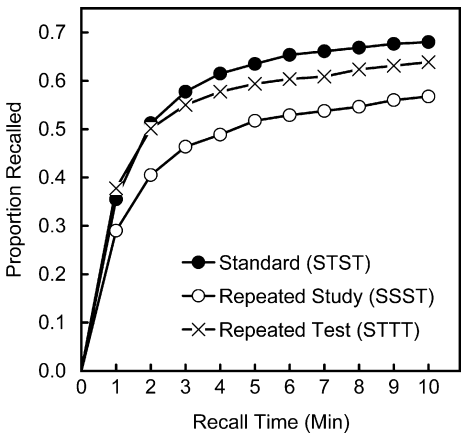

Fig. 3. Cumulative proportion of words recalled on the final test in Experiment 1.

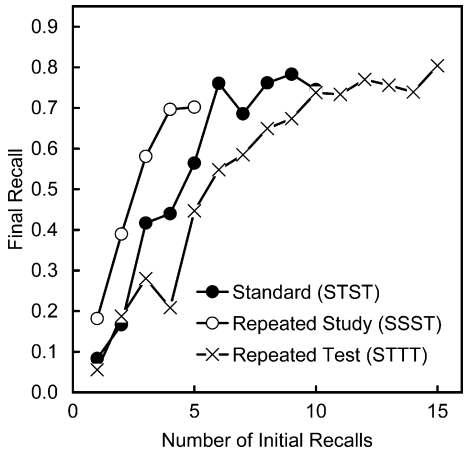

Fig. 4. Conditional analysis showing the probability of final recall of words given the number of times words were recalled initially in the learning phase in Experiment 1.

reveals a consistent pattern for all three conditions, revealing that the likelihood of recalling items on the final test increased as a function of the number of times the items were recalled during the initial learning phase. This relationship between the number of times items were recalled initially in the learning phase and recall on the final test holds true for all three of the conditions; even within the repeated study (SSST) condition there is a dramatic effect of repeated recall, such that recalling items one time in the learning phase led to about 20% recall on the final test whereas repeatedly recalling items 4 or 5 times led to about 70% recall on the final test.

Discussion

In Experiment 1, even though subjects in the repeated study (SSST) condition studied the entire list 15 times in the learning phase and those in the repeated test (STTT) condition studied it only 5 times, the tested group recalled more on the final test 1 week later, clearly showing that repeated testing enhanced long-term retention. Experiment 1 was inspired by Tulving's (1967) work demonstrating that learning occurs during test trials, and we replicated Tulving's basic point that the learning curves for these three conditions are remarkably similar, though we did find an advantage of the standard study-test condition. Most notably, increasing the number of study trials did not improve learning and actually produced the worst retention 1 week later. We suspect that had we employed a pure study condition in which subjects studied the list 20 times and never took a test, recall a week later would have been much worse than in the current repeated study condition (see Hogan & Kintsch, 1971).

Although Tulving (1967) and others (e.g., Lachman & Laughery, 1968) implied that a test trial produces learning equivalent to a study trial, our results show that tests produce more learning than study trials in that repeated testing improved long-term retention relative to repeated studying, a result that conceptually replicates other findings (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Wheeler et al., 2003). The implication of this result is that tests not only assess learning but also greatly enhance it, promoting long-term retention. However, it appears that alternating study and test trials (the standard condition) may represent an optimal condition for enhancing learning, at least relative to the other two schedules we employed. The standard condition may be best because this condition involves more frequent feedback than the other conditions, in the form of more frequent test trials followed by study trials, which may serve as a kind of feedback. That is, following a test trial subjects may be able to recognize items they did not recall in the subsequent study phase, encode those items well, and then recall them first on the next test trial (Battig et al., 1965). In addition, as shown in Fig. 2, having three test trials potentiates learning on the next study trial to an even greater extent (Izawa, 1971). Experiment 2 was carried out as a further investigation of how repeated studying and repeated testing affect long-term retention.

Experiment 2

In Experiment 2 we investigated what kinds of repeated practice lead to superior long-term retention. The results obtained in Experiment 1 indicated that alternating study and test trials led to the best initial learning and later retention. As discussed above, the standard condition may be better than the other two because it provides feedback soon after testing. In the STST condition used in Experiment 1, nine of the test trials were followed immediately by a study trial, whereas in the other two conditions only four of the test trials were followed immediately by a study trial. The advantage of the standard condition could also arise because study and test trials are spaced throughout the learning phase in this condition, whereas the other conditions involved massed repeated studying or repeated testing, and it is well known that spaced practice leads to gains in long-term retention (e.g., Glenberg, 1976; Melton, 1970; for a review, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

Experiment 2 investigated the effects of four different learning conditions on long-term retention. In one condition the subjects studied and recalled the entire list during alternating study and test trials, the standard STST condition used in Experiment 1. In another condition, the subjects studied the list in two consecutive study trials and then recalled in two consecutive test trials (SSTT). These two conditions involve the same

number of study and test trials as in the standard condition, but the distribution of study and test trials differs. Based on the idea that tests facilitate studying on a following study trial (and also based on the positive effects of spaced practice) we expected the STST condition to outperform the SSTT condition, despite their nominal similarity in the number of study and test trials. We can find no prior free recall experiment that provides this contrast. However, the issue of which condition will be superior is in some doubt, because a countervailing factor that could boost performance in an SSTT condition is the potentiating effects of having two test trials prior to a study trial rather than one (Izawa, 1971).

We also investigated how different types of repeated studying or testing affected learning and retention by using two "dropout" conditions in which items were dropped from further studying or testing depending on whether they had been recalled on prior tests. In one condition, subjects studied the entire list of words during a study trial, recalled as many of the words as they could during a test trial, and then restudied only the words that they had not recalled on the previous test trial, and then were tested on the whole list; the procedure was repeated several times. We denote this condition as STS_NT, where S_N indicates that subjects restudied only non-recalled items; once an item was recalled, it was dropped from further study. Note that subjects still attempted to recall the entire list on every test trial despite the shorter study list each time. Bahrick (1979) and Thompson et al. (1978) have used similar conditions, and the procedure is also similar to the selective reminding procedure developed by Buschke (1973). In another dropout condition, subjects restudied the words they had not recalled on the previous test (as in the previous condition), but were told to recall only the words that they had studied on the previous study trial (denoted STS_NT_N, where T_N indicates that subjects recalled only non-recalled items). In this condition, each word was recalled only one time before it was dropped from the study and test phases, and the number of items studied and tested grew smaller over trials. This dropout condition is similar to what study guides often instruct students to do in when studying facts by using flash cards and other methods: Drop material that is already "learned" (or recallable) from further practice and focus on material that is not yet learned. We test the efficacy of this procedure, relative to three other procedures, on long-term retention measured a week later.

Finally, we made a few minor procedural changes from Experiment 1 worth noting. First, to increase the difficulty of the task in Experiment 2, we increased the length of the study list to 60 words and reduced the duration of the learning phase to 4 cycles (thus the learning phase lasted 16 trials). Second, because the Tulving-Colotla procedure could not be easily applied to the dropout conditions, we had subjects perform a 30 s distracter task after each study trial to eliminate primary memory effects (Glanzer & Cunitz, 1966). Finally, in Experiment 2 the subjects typed their responses into the computer during recall trials rather than writing them by hand. After the subjects had entered each recall response, their response was displayed in a list on the computer screen, in order to make this computerized free recall task as similar to written recall as possible and to eliminate output monitoring problems that would arise if subjects could not see the words they had already recalled on the test (see Gardiner, Passmore, Herriot, & Klee, 1977). Otherwise, the procedure was the same as that used in Experiment 1. Subjects studied and recalled a list of words across several trials under one of four conditions (STST, SSTT, STS_NT, or STS_NT_N) during an initial learning phase. The subjects then took a final free recall test after a 1 week retention interval.

Methods

Subjects

Sixty Washington University undergraduates, ages 18–24, participated in exchange for course credit. None of the subjects had participated in Experiment 1.

Materials

Sixty unrelated words were selected from the norms of Paivio et al. (1968). Twenty additional medium frequency words were added to the set of 40 words used in Experiment 1, using the same selection criteria.

Design

Fifteen subjects were assigned to each of the four learning conditions (STST, SSTT, STS_NT, and STS_NT_N). In the STST condition, subjects studied and recalled the list of words during alternating study and test trials. In the SSTT condition, subjects studied during the list on two consecutive study trials (and performed the arithmetic task after each study trial) and then recalled the list during two consecutive test trials. In the other two learning conditions, items were dropped from further studying or testing once they had been previously recalled. In the STS_NT condition, subjects studied and recalled the words during alternating trials, but words that they recalled on a test were dropped from the next study trial. Subjects were still instructed to recall all of the words on each test trial. Likewise, in the STS_NT_N condition, words that had been recalled were dropped from the next study trial, but in this condition the subjects were told that on each test they needed to recall only the words that they studied on the previous study trial (thus, they would recall all 60 words one time in the learning phase). The learning phase involved 4 cycles, so subjects studied and recalled the words during a total of 16 trials. Subjects returned to the lab a week later for a final free recall test.

Procedure

The procedure was similar to that used in Experiment 1. Subjects were tested in groups of five or fewer. At the beginning of each study trial, a "Ready" prompt was shown on the computer screen for 1 s, and then the 60 words were presented on the screen at a rate of 2 s per word, in a different random order on each study trial. Thus, each study trial lasted 2 min in the STST and SSTT conditions, but the duration of study trials varied in the STS_NT and STS_NT_N conditions depending on how many words were presented. After every study trial, the subjects performed a 30 s distracter task that involved verifying multiplication problems. At the beginning of each test trial, a "Recall" prompt and a cursor appeared on the screen, and the subjects were told to type as many of the words as they could recall, in any order. The subjects were instructed to press the Enter key after they had typed each response, and upon doing so the response they had typed was added to a list of their responses that remained displayed on the computer screen throughout the recall trial. Each test trial lasted 2 min in all four learning conditions (regardless of the number of items to be recalled).

The subjects took a final free recall test 1 week after the learning phase, in which they were given 10 min to recall as many of the words as they could, in any order. Just as they had done during the learning phase, the subjects typed their responses into the computer, and the computer recorded the time associated with each recalled response in order to obtain measures of cumulative recall.

Results and discussion

Learning phase

Because the four learning conditions differed in their recall requirements (the STS_NT_N condition involved recalling only words that had not been previously recalled, while the other conditions involved recalling the entire list of words), we compared the conditions on cumulative recall during the learning phase. That is, Fig. 5 shows the proportion of items subjects had recalled at least once as measured on each test trial, thus holding the four conditions to the same performance criterion. Fig. 6 shows traditional learning curves, the proportion of words recalled on each trial, for the three conditions that required subjects to recall the entire list on each trial. Comparing Figs. 5 and 6 shows that the pattern of results was essentially the same by both scoring methods for these three conditions; thus we restrict our analysis to the cumulative learning curves.

Fig. 5 clearly shows that the STS_NT_N condition showed the fastest initial learning of the list, reaching 99% recall by the end of the learning phase. (One subject in this condition failed to recall one item; thus mean performance was not quite 100%. Recall that a 30 s distractor task occurred after items were presented in all conditions.) The occurrence of mistaken repeated recall in the STS_NT_N condition was infrequent: Only two subjects mistakenly recalled one item more than once. The STST and STS_NT conditions produced similar initial learning (87% and 83% recall by the end of the learning phase), and both conditions showed better learning than the SSTT condition (75%). A 4 (learning condition) × 4 (cycle) ANOVA performed on the cumulative proportion of words recalled on the fourth trial in each cycle revealed a main effect of learning condition, $F(3, 56) = 23.85$, $\eta_p^2 = .56$, a main effect of cycle, $F(3, 168) = 1534.32$, $\eta_p^2 = .97$, and a condition × cycle interaction, $F(9, 168) = 10.21$, $\eta_p^2 = .35$. Although recall

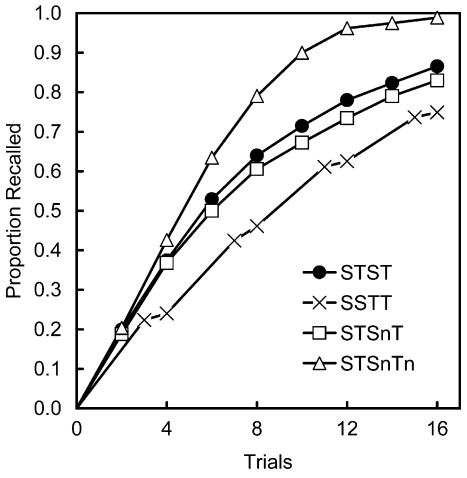

Fig. 5. Cumulative learning (proportion of words recalled for the first time on each trial) during the learning phase in Experiment 2.

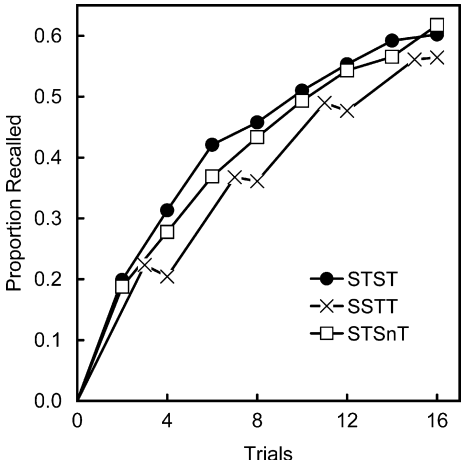

Fig. 6. Proportion of words recalled on each trial in the learning phase (the traditional learning curve) in Experiment 2.

was comparable in the four conditions early in the learning phase, the STS_NT_N condition rapidly diverged from the other conditions and reached ceiling levels of performance.

Final recall

Fig. 7 shows cumulative recall on the final test given 1 week after learning and reveals a striking pattern of results. By the end of the 10 min recall period, the standard STST condition and the STS_NT condition produced equivalent final recall (44%). Thus repeatedly studying items that had been previously recalled (in the STST condition) did not enhance retention relative to dropping those items from further study (in the STS_NT condition). (Alternatively, dropping recalled items from further studying did not improve retention by making repeated studying easier or more effective.) Both the STST and STS_NT conditions outperformed the SSTT condition (36%, ds = .41 and .47, p_{rep}s = .78 and .82, respectively), demonstrating a positive effect of more frequent feedback on retention. But most strikingly, all three conditions just discussed outperformed the STS_NT_N condition which led to 21% recall on the final test, despite the fact that subjects in STS_NT_N constituted the only group that had recalled all of the items once in the learning phase. The STST, STS_NT, and SSTT conditions all outperformed the STS_NT_N condition by a large margin (ds = 1.37, 1.69, and 1.10, p_{rep}s = .99, .99, and .97, respectively).

Converging evidence is shown in Fig. 8, which shows the conditional analysis of the probability of final recall given the number of times items were recalled initially in the learning phase. As in Experiment 1, this figure shows a systematic pattern of results. In the STS_NT_N condition, all items were recalled one time, and final recall in this condition looks the same as final recall of items

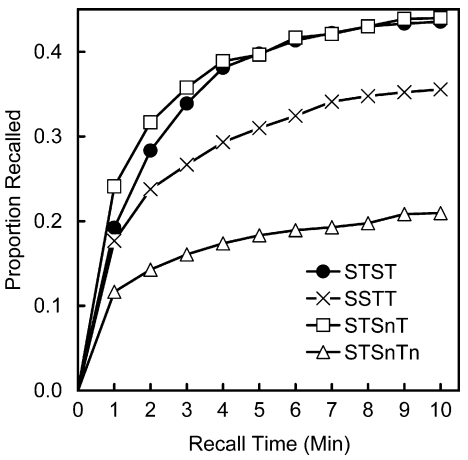

Fig. 7. Cumulative proportion of words recalled on the final test in Experiment 2.

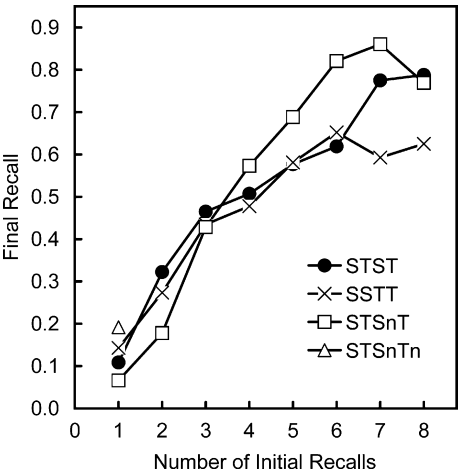

Fig. 8. Conditional analysis showing the probability of final recall of words given the number of times words were recalled initially in the learning phase in Experiment 2.

recalled only once in the other three conditions. However, repeatedly recalling items in the other three conditions increased the likelihood that those items would be recalled on the final test. Thus, a simple procedural change – requiring subjects to repeatedly recall the entire list in the STS_NT condition relative to the STS_NT_N condition – produced a greater than 100% improvement in long-term retention.

General discussion

Tulving (1967) showed that tests not only assess learning but also produce it. His results showed that, within broad limits, a test can substitute for a study trial and produce the same amount of learning. We adopted Tulving's general procedure and replicated his finding that the three conditions used in his experiment (STST, SSST, and STTT) produced similar learning curves. However, we also showed that repeated studying and repeated testing lead to quite disparate results on a later final test given after a delay. A test trial has much more impact than a study trial on long-term retention. In Experiment 1, we showed that repeated testing during learning produced better long-term retention than repeated studying. Even though subjects studied the list 15 times in the study condition (SSST) and they studied it only 5 times in the test condition (STTT), the test condition led to better long-term retention one week later. In Experiment 2, repeatedly studying items that had already been recalled did little to enhance retention, whereas repeated testing of recalled items had large positive effects on long-term retention. Together these results point to our main conclusion: Repeated retrieval is the key to enhancing later retention (see too Roediger, 2000).

Why should additional retrieval practice improve retention while additional encoding practice (beyond some necessary amount) does not? We (Roediger & Karpicke, 2006a, 2006b) have argued that the positive effects of testing on later retention can be partly understood according to the concept of transfer-appropriate processing (see Kolers & Roediger, 1984; Morris, Bransford, & Franks, 1977; Roediger, 1990; Roediger et al., 2002). The idea behind transfer-appropriate processing is that memory performance will benefit to the extent that the processes engaged in during initial learning overlap with the processes required to perform well on a final test. In the case of the testing effect, when subjects take an initial test they are required to engage in retrieval processes to access information stored in memory, and practicing this skill on initial tests will transfer positively and enhance performance on later tests given in the future. In the present experiments, repeatedly studying the list provided additional exposure to the all of the list items whereas repeated testing required subjects to practice retrieval, and thus in the repeated test conditions subjects practiced the skill necessary to recall the list items in the long-term.

Another concept useful for understanding the present results is Bjork's (1994, 1999) hypothesis of introducing desirable difficulties to enhance learning. Bjork has compiled a variety of evidence indicating that techniques that may promote rapid initial learning (such as repeated, massed studying) will often lead to poor long-term retention and, likewise, techniques that make initial learning slower or more effortful often enhance long-term retention. The testing effect is one example of a desirable difficulty: Testing leads to better long-term retention after a delay than does repeated study, even though massed studying often produces a boost shortly after learning (see Roediger & Karpicke, 2006b; Thompson et al., 1978; Wheeler et al., 2003). Experiment 2 shows another example of repeated retrieval as a desirable difficulty. In Experiment 2, eliminating repeated recall in the STS_NT_N condition by requiring subjects to recall each item only once led to rapid acquisition of the entire list, while the other conditions that involved repeatedly recalling the entire list were slowed in cumulative learning performance. However, dropping items from repeated recall in the STS_NT_N condition produced much worse retention on a one week delayed test. Requiring repeated recall of the list made initial learning more difficult but greatly enhanced long-term retention.

There are other possible explanations for why the STS_NT_N condition performed poorly relative to the repeated test conditions (STST and STS_NT) aside from the idea that repeated retrieval practice was responsible for enhancing long-term retention in the repeated test conditions. One idea is that asking subjects only to recall previously non-recalled words disrupted the organizational strategies that they might normally use to accomplish the free recall task. When subjects did not repeatedly recall the entire list, they could not organize their recall output in the same way as when they repeatedly recalled the entire list. Another idea is that instructing subjects to recall only words that they studied on the previous study trial, and telling them not to recall words they had recalled on previous test trials, acted like a directed forgetting instruction and subjects intentionally forgot words once they had recalled them. Recalling only non-recalled words might also have produced some other type of retrieval interference, such as retrieval-induced forgetting of words that had already been recalled. These alternative explanations deserve consideration and await future research, but for the time being we can express some doubt that they explain the entire effect. In other research using paired associate materials, we replicated the effect of repeatedly testing the entire set of materials vs. dropping recalled words from further testing. We found similar large effects of repeated retrieval practice on long-term retention relative to dropping pairs once they were recalled. Because we used word pairs, we did not instruct subjects not to recall previously recalled words, and because there is no reason to assume competition among unrelated word pairs, retrieval-induced forgetting of non-tested pairs is unlikely in this case. Although these alternative explanations are interesting in their own right, we believe that repeated retrieval practice is responsible for the advantage of the repeated test conditions (STST and STS_NT) over the STS_NT_N condition.

Our results suggest that continuing practice after material has been learned well enough to be recalled, or overlearning, can be effective for enhancing long-term retention (Postman, 1962). Overlearning is sometimes recommended as a technique for improving learning in education and training, and one meta-analysis concluded that studies do generally show positive effects of overlearning on long-term retention (Driskell, Willis, & Copper, 1992). However, Rohrer, Taylor, Pashler, Wixted, and Cepeda (2005) recently argued that overlearning may not benefit retention at very long retention intervals (e.g., 9 weeks), and thus the effectiveness of overlearning at long delays is currently an open question. Our results indicate that an important distinction should be made in evaluating the effectiveness of overlearning. While additional studying or encoding practice had little or no effect on retention, repeated testing or retrieval practice had profound effects on long-term retention. The effectiveness of overlearning probably depends on the type of practice involved in the particular overlearning procedure used.

Recently, researchers have made increasing efforts toward investigating how students allocate their study time while they are learning. This research indicates that when students are asked to assess how well they have learned particular items (judgments of learning) and

then are asked to select items to study again (or the amount of time they spend restudying particular items is measured), students will allocate more study time to items given lower judgments of learning (e.g., Nelson & Leonesio, 1988; Nelson, Dunlosky, Graf, & Narens, 1994; Mazzoni & Cornoldi, 1993), although some exceptions to this general pattern exist (e.g., Metcalfe & Kornell, 2003, 2005; Thiede & Dunlosky, 1999). The assumption guiding research on study-time allocation is the same assumption discussed in the introduction that is widely held by many students and educators: Learning happens during studying, and therefore effective study-time allocation is critical for optimizing learning. We believe that the results of our experiments suggest that how students allocate their study time may not always have much to do with how well they remember material in the long-term. The idea of effective study-time allocation suggests that material that is learned or recallable should be dropped from study and further studying should be allocated to material that is not yet learned or recallable. Yet in our Experiment 2, when recalled items were dropped from study and only non-recalled items were restudied (representing effective study-time allocation) this step made absolutely no difference for later retention of the material. Furthermore, dropping recalled items from further practice altogether (in the STS_NT_N condition) led to the worst long-term retention. We are not alone in expressing doubt about the importance of study-time allocation. Metcalfe and Kornell (2005) recently cautioned that effective study-time allocation may not necessarily promote effective learning: "We still do not know whether what [students do when allocating study-time] enhances their learning, or is in any way optimal. Until we have answered the still-open question of efficacy, despite the subtlety of people's strategies ... we cannot fully endorse the idea that they are exerting *good* metacognitive control" (p. 476, italics in original).

Testing is a powerful means of improving learning and long-term retention. The practical implication of our results is that students should test themselves repeatedly while they are learning, not just because self-testing provides knowledge of results that can guide future studying, but also because the act of retrieving information leads to large benefits for retention. However, we doubt that many students test themselves while they are studying and, if they do, students most likely use testing as a means of generating feedback about whether or not material is learned rather than using the act of retrieval itself as a method of promoting learning. If students only use self-testing to assess their knowledge and then allocate their studying accordingly, they would drop "learned" material from further studying and testing, because additional tests would not provide the student with any other information about whether the item is recallable. Even though repeated retrieval practice is a powerful way to enhance long-term retention, it is not clear that students or educators view the act of taking a test as a learning device or use testing as a tool to enhance learning. Future research in this area has the potential to inform the kinds of study techniques recommended to students and educators.

References

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

Battig, W. F., Allen, M., & Jensen, A. R. (1965). Priority of free recall of newly learned items. *Journal of Verbal Learning and Verbal Behavior, 4*, 175–179.

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 peformance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 9*, 689–698.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior, 12*, 543–550.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

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

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615–622.

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

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *Journal of Gerontology: Psychological Sciences B, 52*, P178–P186.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.) New York: Dover (original work published 1885).

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16*, 45–54.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*, 40.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*, 351–360.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1–16.

Greene, R. L. (1992). *Human memory: Paradigms and paradoxes*. Hillsdale, NJ: Erlbaum.

Hogan, R. M., & 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. (1971). The test-trial potentiating model. *Journal of Mathematical Psychology, 8*, 200–224.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*, 345–353.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior, 23*, 425–449.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162.

Lachman, R., & Laughery, K. R. (1968). Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology, 76*, 40–50.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study-time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122*, 47–60.

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

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.

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.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

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.

Paivio, A., Yuille, J. C., & Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs, 76*(1), 1–25.

Patterson, K. E. (1972). Some characteristics of retrieval limitation in long-term memory. *Journal of Verbal Learning and Verbal Behavior, 11*, 685–691.

Postman, L. (1962). Retention as a function of degree of overlearning. *Science, 135*, 666–667.

Rock, I. (1957). The role of repetition in associative learning. *American Journal of Psychology, 70*, 186–193.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043–1056.

Roediger, H. L. (2000). Why retrieval is the key process to understanding human memory. In E. Tulving (Ed.), *Memory, consciousness, and the brain: The Tallinn conference* (pp. 52–75). Philadelphia, PA: Psychology Press.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory, 10*, 319–332.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Roediger, H. L., & Thorpe, L. A. (1978). The role of recall time in producing hypermnesia. *Memory & Cognition, 6*, 296–305.

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

Rosner, S. R. (1970). The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior, 9*, 69–74.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self regulated study: An analysis of selection of items for study and self paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024–1037.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210–221.

Tulving, E. (1967). The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184.

Tulving, E. (1968). Theoretical issues in free recall. In T. Dixon & D. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 2–36). Englewood Cliffs, NJ: Prentice-Hall.

Tulving, E., & Colotla, V. A. (1970). Free recall of trilingual lists. *Cognitive Psychology, 1*, 86–98.

Watkins, M. J. (1974). The concept and measurement of primary memory. *Psychological Bulletin, 81*, 695–711.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580.