Journal of Experimental Psychology:
Learning, Memory, and Cognition
2012, Vol. 38, No. 3, 737–746

0278-7393/11/$12.00 DOI: 10.1037/a0026166

Why Is Test–Restudy Practice Beneficial for Memory? An Evaluation of the Mediator Shift Hypothesis

Mary A. Pyc and Katherine A. Rawson
Kent State University

Although the memorial benefits of testing are well established empirically, the mechanisms underlying this benefit are not well understood. The authors evaluated the mediator shift hypothesis, which states that test–restudy practice is beneficial for memory because retrieval failures during practice allow individuals to evaluate the effectiveness of mediators and to shift from less effective to more effective mediators. Across a series of experiments, participants used a keyword encoding strategy to learn word pairs with test–restudy practice or restudy only. Robust testing effects were obtained in all experiments, and results supported predictions of the mediator shift hypothesis. First, a greater proportion of keyword shifts occurred during test–restudy practice versus restudy practice. Second, a greater proportion of keyword shifts occurred after retrieval failure trials versus retrieval success trials during test–restudy practice. Third, a greater proportion of keywords were recalled on a final keyword recall test after test–restudy versus restudy practice.

Keywords: testing effects, retrieval practice, mediator use, mediator shift hypothesis

In the memory literature, tests have traditionally been used as a means of assessment. However, researchers have discovered that tests can be used to promote memory (i.e., *testing effects*; for reviews, see Rawson & Dunlosky, 2011; Roediger & Butler, 2011). Testing effects have been demonstrated across a wide range of materials, test formats, and learners. For example, testing effects have been found for foreign language word pairs (e.g., Pyc & Rawson, 2007), expository text (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008), word lists (e.g., Carpenter & DeLosh, 2006), general knowledge facts (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), face-name pairs (e.g., Landauer & Bjork, 1978), and visuospatial materials such as map learning (e.g., Carpenter & Pashler, 2007). Testing effects have been demonstrated across a variety of test formats including cued recall tests (e.g., Pyc & Rawson, 2009), free recall tests (e.g., Karpicke & Roediger, 2007), and multiple-choice tests (e.g., Butler, Marsh, Goode, & Roediger, 2006). Finally, testing effects have been demonstrated across the lifespan including preschool-aged children (e.g., Fritz, Morris, Nolan, & Singleton, 2007), young adults (e.g., Karpicke & Roediger, 2008), and older adults (e.g., Logan & Balota, 2008).

Researchers have discussed two different effects that testing may have on learning and retention, referred to as *direct* and *mediated* effects (Roediger & Karpicke, 2006). The direct effect of testing refers to facilitation in learning that arises from the act of taking the test itself. To evaluate the direct effects of testing, researchers typically compare final test performance after practice tests only (with no restudy opportunities) versus after restudy only or after no practice. The modal finding is that test-only outperforms the other two.

However, research has also shown that testing followed by restudy is more effective than either testing or restudying alone, suggesting effects of testing above and beyond those from the act of testing itself (e.g., Cull, 2000). Roediger and Karpicke (2006) referred to these as mediated effects of testing, or facilitation in learning that arises from the influence of testing on subsequent encoding during restudy. For example, testing may lead an individual to change from a less effective to a more effective encoding strategy on a subsequent restudy trial, leading to enhanced retention of that information. To evaluate the mediated effects of testing, researchers typically compare performance on a final retention test between individuals who receive test–restudy trials during practice with individuals who receive only restudy trials during practice. The modal finding is that individuals engaged in test–restudy practice outperform those engaged in restudy only. Although much of the early research on testing effects focused on direct effects of testing in the absence of restudy, an increasing amount of recent research has examined the effects of testing with restudy. For example, Rawson and Dunlosky (2011) summarized the methods used in 161 experiments published from 2000–2010, and 61% included one or more test–restudy conditions. To foreshadow, the effects of test–restudy practice are also the focus of the current research.

Although testing effects are empirically well established, *why* testing is beneficial for memory is not well understood. That is, very few studies have systematically evaluated theoretical explanations for the memorial benefits of testing. Furthermore, existing

This article was published Online First November 7, 2011.

Mary A. Pyc and Katherine A. Rawson, Department of Psychology, Kent State University.

The research reported here was supported by supported by a Collaborative Award to Katherine A. Rawson from the James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior.

Correspondence concerning this article should be addressed to Mary A. Pyc, who is now at Department of Psychology, Box 1125, Washington University in St. Louis, St. Louis, MO 63130. E-mail: mpyc@wustl.edu

theories have primarily focused on explaining direct effects of testing, whereas mediated effects are even less well understood. Accordingly, the primary goal of the current work was to test a theoretical account of why test–restudy practice promotes memory. Below we discuss two theoretical accounts, the elaborative retrieval hypothesis and the retrieval effort hypothesis, that have been proposed to explain the direct effects of testing. Our primary purpose in describing these accounts is to contrast these theories with the one evaluated in the current experiments, although we note that these accounts are not mutually exclusive, as effects can arise from multiple sources.

Theoretical Accounts of Testing Effects

The *elaborative retrieval hypothesis* (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006) is based on the assumption that performance will be better on a final retention test when more elaborative versus less elaborative memory traces are formed during practice (cf. encoding variability hypothesis, McDaniel & Masson, 1985). Retrieval practice leads to the formation of elaborative memory traces because searching through memory for the correct answer activates other related information, and connections are formed between the cue, the target answer, and other related information in memory. For restudy trials, however, this same search through memory is not necessary because the target information is explicitly presented. Therefore, more elaborative connections are formed after test versus restudy practice, which presumably improves later memory on a final retention test.

Another theoretical account of the direct effects of testing comes from the *desirable difficulty framework* (Bjork, 1994). One basic claim of this framework is that performance on a retention test will be greater when processing during encoding is successful but difficult compared with when processing during encoding is successful but less difficult. Pyc and Rawson (2009) recently forwarded a more specific instantiation of the desirable difficulty framework to account for the effects of testing on memory. According to their *retrieval effort hypothesis,* performance on a retention test will be greater when correct retrievals during practice are more difficult compared with when correct retrievals during practice are less difficult.

Both of the accounts discussed above primarily have to do with the benefit of retrieval itself, particularly when retrieval is successful during encoding. In contrast, the account we propose below has to do with how retrieval attempts influence what happens during restudy, particularly after a retrieval failure during encoding. Specifically, the *mediator shift hypothesis* states that individuals will shift to using more effective mediators within a given strategy as a function of the status of retrieval (i.e., retrieval failure or retrieval success) during practice. According to the mediator shift hypothesis, the memorial benefits of test–restudy practice are greater than the memorial benefits of restudy practice because individuals monitor and modify mediators when they cannot correctly recall an item during test–restudy practice, whereas individuals who only restudy are less likely to shift mediators because they do not experience retrieval failure.

The mediator shift hypothesis is adapted from the *strategy shift hypothesis* (Bahrick & Hall, 2005), which was originally proposed to account for spacing effects (i.e., the memorial benefits of distributed versus massed practice), but it may also be relevant to explaining why performance is greater on retention tests after test–restudy practice compared with restudy practice. The basic claim of the strategy shift hypothesis is that retrieving information from memory during practice enables individuals to evaluate the effectiveness of strategies used to encode information. More specifically, when individuals experience retrieval failure, they evaluate the effectiveness of the strategy they used to encode an item and shift from a less effective to a more effective strategy during subsequent restudy, which is beneficial for later recall. With respect to spacing effects, their claim was that learners experience more retrieval failures when practicing items over longer versus shorter intervals, and thus, learners are more likely to shift to effective strategies with distributed versus massed practice. Likewise, the same logic can be applied to explaining testing effects. Whereas learners experience retrieval failures during test–restudy practice, individuals engaged in restudy only do not experience retrieval failures and thus will be less likely to shift to using effective strategies.

The primary goal of Bahrick and Hall (2005) was to evaluate the strategy shift hypothesis for the spacing effect, rather than the memorial benefits of test–restudy practice versus restudy practice. Nonetheless, we briefly describe their studies here because their methodology provided a foundation for the current research. Participants received instructions explaining various encoding strategies they could use to learn Swahili-English word pairs, including rote repetition, verbal elaboration (i.e., generating a word or sentence connecting the cue and target), and visual elaboration (i.e., forming an image connecting the cue and target). Participants learned word pairs across four practice sessions that were either massed (each session began when the previous session ended) or distributed (practice sessions spaced across days). After an initial study trial with each word pair, participants had a strategy trial in which they reported which kind of strategy they used to encode the word pair. After the strategy report, participants had a self-paced test trial in which they were given the cue and had to retrieve the target. Items that were correctly recalled were dropped from practice, whereas items that were incorrectly recalled received another restudy trial. After the restudy trial, participants again reported the kind of strategy they used to encode the word pair. This process continued until all items were correctly recalled once. The next practice session occurred immediately for participants assigned to massed practice, whereas participants assigned to distributed practice came back for their next session either 1 day later or 14 days later. Each participant had four practice sessions. Participants came back for a final cued-recall test 14 days after their last practice session.

On the final retention test, performance was greater after distributed practice compared with massed practice. Importantly, strategy reports showed that participants more often used visual or verbal elaboration during distributed practice compared with massed practice. Presumably, participants did not experience as many retrieval failures during massed versus distributed practice and thus did not have the opportunity to evaluate the effectiveness of their encoding strategies. A second experiment did not collect strategy reports, so it is difficult to determine which strategies participants used to encode items (if they used any). However, participants took longer restudying items after retrieval failure trials compared with retrieval success trials, which the authors took

as evidence that participants were shifting encoding strategies after retrieval failure trials during practice.

Results from Bahrick and Hall (2005) provided some support for the strategy shift hypothesis for the spacing effect. However, their experiments had some limitations that are discussed here to highlight important changes for the current experiments. First, Bahrick and Hall evaluated shifting of global strategies (e.g., from rote repetition to imagery). However, participants may also have shifted specific mediators within a given strategy. For example, participants may have shifted from a less effective visual elaboration to a more effective visual elaboration after retrieval failure trials during practice (e.g., shifting from a noninteractive image to an interactive image). A second limitation of Bahrick and Hall (2005) is that Experiment 2 used latencies after retrieval failure trials, versus retrieval success trials, to infer that participants were shifting strategies after retrieval failure trials during practice, but they did not collect strategy reports to confirm this assumption. Therefore, it is difficult to determine the extent to which participants were using a longer amount of time restudying items after retrieval failure trials, because they were generating new strategies or if they were engaged in some other activity. Finally, because Bahrick and Hall's (2005) goal was to evaluate the strategy shift hypothesis for the spacing effect, their experiments were not designed to compare test–restudy and restudy schedules of practice. Therefore, the current experiments compared shifting of mediators during test–restudy practice versus during restudy practice.

Current Experiments

In the current experiments, we evaluated the mediator shift hypothesis using a task in which individuals were asked to learn foreign language word pairs. Therefore, it was important to instruct individuals to use a global strategy that is amenable to learning foreign language word pairs. One effective strategy for learning these word pairs is a keyword encoding strategy (e.g., Fritz, Morris, Acton, Voelkel, & Etkind, 2007; McDaniel, Pressley, & Dunay, 1987; Raugh & Atkinson, 1975). In the current experiments, the keyword strategy involved two basic steps. First, learners are instructed to generate an English word that is orthographically or phonologically similar to the foreign language cue. Second, learners are instructed to relate the generated English keyword to the English target so that the two words are semantically connected. For example, consider the Swahili-English word pair *wingu–cloud*. An English word that is similar to *wingu* is "wing," and the connection between *wingu* and *cloud* could be "birds have wings and fly in clouds." Therefore, when prompted with *wingu*, a learner would first retrieve the keyword "wing" and then relate it to birds, which fly in clouds, to retrieve the target *cloud*.

The basic method for the current experiments is described below to facilitate understanding of the predictions of the mediator shift hypothesis. First, participants received an initial study trial followed by a keyword report trial for each item. For the initial study trials both the cue and target were presented, and participants were told to use the keyword encoding strategy to help them remember the word pairs. For keyword report trials, participants were asked to type in the keyword they used to associate the cue and target. Second, items were presented for either test–restudy practice or restudy practice. For test trials, the cue (e.g., *wingu*) appeared on the computer screen, and participants typed in the target answer (e.g., *cloud*). For restudy trials, both the cue and target words appeared on the computer screen. After each restudy trial in both groups, participants completed a keyword report trial in which they reported the specific keyword used to encode a given item. Third, participants returned 2 days later for a final cued-recall test, followed by a test of their memory for the keyword mediators they generated during Session 1.

For all experiments, final cued-recall performance was predicted to be greater for items learned with test–restudy practice versus restudy practice. The mediator shift hypothesis makes three basic predictions regarding why final test performance will be greater after test–restudy practice versus restudy practice. Prediction 1 is related to keywords generated during practice and states that a greater proportion of keyword shifts (i.e., changing from a keyword that was previously generated for a given word pair to a new keyword for that word pair) will occur during test–restudy practice versus restudy practice. This prediction is based on the assumption that individuals engaged in test–restudy practice will experience retrieval failure during practice. These individuals will be better able to evaluate the effectiveness of keywords (compared with individuals engaged in restudy practice) and will shift to effective keywords more often than will individuals engaged in only restudy practice.

Prediction 2 specifically relates to keyword use during test–restudy practice. Prediction 2 is that a greater proportion of keyword shifts will occur after retrieval failure trials versus retrieval success trials during test–restudy practice. On retrieval failure trials during practice, individuals presumably can evaluate the effectiveness of the keyword they generated for a given item and will shift from a less effective keyword to a more effective keyword.

Prediction 3 is related to keyword recall after the final test. Prediction 3 is that the proportion of keywords generated during practice that are recalled on the final keyword recall test will be greater for items learned with test–restudy practice compared with items learned with restudy practice. The premise of the mediator shift hypothesis is that testing during practice leads to the generation of more effective mediators. Two factors have been shown to influence the effectiveness of mediators: mediator retrieval and mediator decoding (e.g., Dunlosky, Hertzog, & Powell-Moman, 2005; Pyc & Rawson, 2010). For a mediator to be effective, it must be recallable when prompted with a cue, and it must be decoded to elicit the target from memory. The final keyword recall test in the current research provides a measure of mediator retrieval, one of the key components of mediator effectiveness.

Experiments 1a and 1b evaluated these three predictions of the mediator shift hypothesis. Experiments 1a and 1b differed only in the number of items learned and the number of practice trials each item received. These experiments are reported together below because they produced similar results. To foreshadow, results from these first two experiments were in line with Predictions 1 and 2 of the mediator shift hypothesis, but failed to support Prediction 3. To provide a stronger test of the mediator shift hypothesis, Experiment 2 was designed to examine more difficult learning conditions, which were predicted to increase the proportion of retrieval failures during practice. Importantly, Experiment 2 was also designed to provide a better test of Prediction 3 by revising the format of the final keyword recall test.

Experiments 1a and 1b

Method

Participants and design. Participants included 53 Kent State University undergraduates in Experiment 1a and 50 in Experiment 1b who participated in return for course credit. Participants were randomly assigned to test–restudy or restudy practice groups, with 23 to 27 participants in each group.

Materials. Items included 30 (Experiment 1a) and 45 (Experiment 1b) Swahili-English word pairs previously normed for item difficulty (Nelson & Dunlosky, 1994). We selected items based on results from a pilot study in which participants were asked to generate keywords for the full set of 100 normed Swahili-English word pairs. The word pairs selected for the current experiments were those for which participants generated the most keywords in the pilot study. These word pairs were selected to increase the likelihood that participants would be able to come up with a keyword for each word pair during practice and to ensure that it was possible to come up with more than one keyword for each word pair.

Procedure. Task instructions and item presentation were administered via computer. All trials were self-paced. On the instruction screen, participants received detailed instructions regarding the keyword encoding strategy, which they were told they should use during the task. The instructions were as follows:

> Your primary task today is to learn the English translation for Swahili words (e.g., *Mshoni-Tailor*). Previous research has indicated that the keyword method is an effective way to promote memory for word pairs such as the Swahili-English pairs you will learn today (e.g., *Mshoni-Tailor*). The keyword method involves generating a word to associate a foreign language word (e.g., *Mshoni*) with an English word that you already know that will help you remember the correct English translation for the Swahili word (e.g., *SHOE* for *MSHONI* because they sound similar. The word *SHOE* is then related to *TAILOR* because at one time *TAILORS* made *SHOES*).

Participants were provided two additional examples of the keyword encoding strategy, which was followed by instructions regarding the remainder of the task (i.e., that they would have multiple restudy or test–restudy trials with each item, depending on the group to which they were assigned). After reading the instructions, participants were prompted to see the experimenter. At this time, the experimenter had participants describe what they would be doing in the task to ensure that all participants fully understood the keyword encoding strategy and the task that they would be completing.

Each word pair was then presented for an initial study trial and an initial keyword report trial. For initial study, the Swahili-English word pair was presented on the screen until the participant pressed a button indicating that they were done studying the item. After the initial study trial for a given item, participants immediately completed a keyword report trial, in which they were asked to report the specific keyword mediator they used to associate the Swahili-English word pair. Upon completing initial study and initial keyword report trials for all items, items received five (Experiment 1a) or three (Experiment 1b) blocks of distributed practice trials involving either test–restudy or restudy only. This distributed practice involved each next practice trial, with a given item being separated by a practice trial with all other to-be-learned items. For the restudy groups, the Swahili-English word pair was presented on the screen until the participant pressed a button indicating that they were done studying. After each restudy trial with a given word pair, participants had a keyword report trial. For the test–restudy groups, participants had a self-paced test trial before each restudy trial. For test trials, only the Swahili word was presented, and participants typed the English translation in a field provided on the screen.

Upon completion of the experimental task, participants were dismissed and reminded to return 2 days later. The second session began with a final cued-recall test. On each final test trial, participants were prompted with a Swahili cue and asked to report the English target. Upon completion of the cued-recall test, participants completed a keyword recall test for all items. For Experiment 1a, participants were prompted with each Swahili-English word pair and asked to recall all of the keywords from Session 1 that were used for each word pair. For Experiment 1b, participants were prompted with the Swahili-English word pair but were given separate response boxes for each trial they had during practice. For both experiments, if participants could not remember the keyword they used for a given word pair, they were instructed to type "I cannot remember the keyword" in the keyword response box.

Results and Discussion

The mean proportion of items correctly recalled on the final test as a function of practice group is reported in Figure 1 for Experiments 1a and 1b. As expected, performance was greater for items learned with test–restudy practice versus restudy practice. Independent samples t tests revealed significant effects of practice group, $t(51) = 3.51$, $p = .001$ (Experiment 1a) and $t(48) = 3.02$, $p = .004$ (Experiment 1b). Thus, both experiments replicated well-established testing effects, which permits evaluation of predictions of the mediator shift hypothesis.

Keyword shifts during practice. To revisit, Prediction 1 is that a greater proportion of keyword shifts will occur during test–restudy practice versus restudy practice. To test this prediction, analyses were conducted on practice trials that met two criteria reflecting the logically necessary conditions for examining keyword shifting: (a) a keyword had been generated for that item on an earlier practice trial, and (b) a keyword was generated on the current practice trial. Keyword shifts were operationalized as the

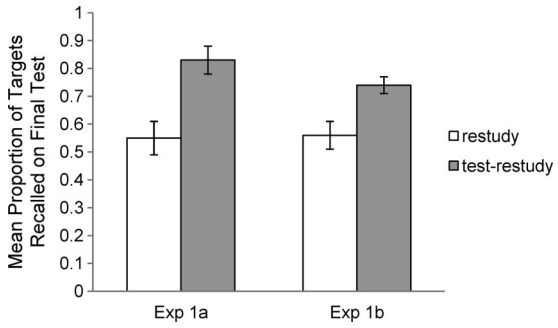

Figure 1. Mean proportion of targets correctly recalled on the final cued recall test as a function of practice group in Experiments 1a and 1b. Error bars represent standard errors.

mean proportion of these trials in which the keyword generated on the current practice trial was different from the keyword generated on the previous trial during practice.

The mean proportion of keyword shifts during practice is reported in Figure 2 for Experiments 1a and 1b. Supporting Prediction 1, a greater proportion of keyword shifts occurred during test–restudy practice versus restudy practice, $t(51) = 2.61$, $p = .012$ (Experiment 1a) and $t(48) = 1.98$, $p = .05$ (Experiment 1b).

Prediction 2 is that a greater proportion of keyword shifts will occur after retrieval failure trials compared with retrieval success trials during test–restudy practice. The proportion of keyword shifts for retrieval failure trials and retrieval success trials in Experiments 1a and 1b is reported in Figure 3. Consistent with Prediction 2, keyword shifts were more likely to occur after retrieval failure trials versus retrieval success trials, $t(26) = 4.56$, $p < .001$ (Experiment 1a) and $t(26) = 5.04$, $p < .001$ (Experiment 1b).

Keyword recall test. Prediction 3 is that the proportion of keywords from Session 1 that are recalled on the keyword recall test in Session 2 will be greater for items learned with test–restudy practice versus restudy practice. Contrary to this prediction, the mean proportion of keywords recalled was similar for test–restudy and restudy groups in Experiment 1a (.88, $SE = .02$ vs. .89, $SE = .03$, respectively) and Experiment 1b (.75, $SE = .03$ vs. .83, $SE = .03$, respectively). An independent samples t test showed no significant difference between restudy and test–restudy groups in Experiment 1a, $t(51) = .30$, $p = .767$. The difference in Experiment 1b, $t(48) = 2.10$, $p = .04$, was in the opposite direction, with a greater proportion of keywords reported on the final keyword recall test in the restudy group compared with the test–restudy group.

One potential explanation for why keyword recall on the keyword recall test was not in line with Prediction 3 concerns the format of the final keyword recall test. Specifically, participants received both the Swahili cue and English target and were asked to recall the keyword they used for the word pair during Session 1. Therefore, participants may have been able to reconstruct the keyword from Session 1 instead of retrieving it from memory. Thus, one important goal of Experiment 2 was to better evaluate Prediction 3 by using cue-only prompts for the keyword recall test.

Taken together, results from Experiments 1a and 1b confirmed Prediction 1 and Prediction 2 of the mediator shift hypothesis. However, one potential limitation of the current experiments is related to the mean proportion of retrieval failures for each word pair during practice. The mean proportion of retrieval failures was relatively low for Experiment 1a (M proportion = .21, $SE = .04$, or about 1 out of 5 trials) and Experiment 1b ($M = .31$, $SE = .03$, or about 1 out of 3 trials), which could limit the extent of keyword shifting during practice. Therefore, one goal of Experiment 2 was to examine more difficult learning conditions, which were predicted to increase the proportion of retrieval failures for each item during practice for the test–restudy group.

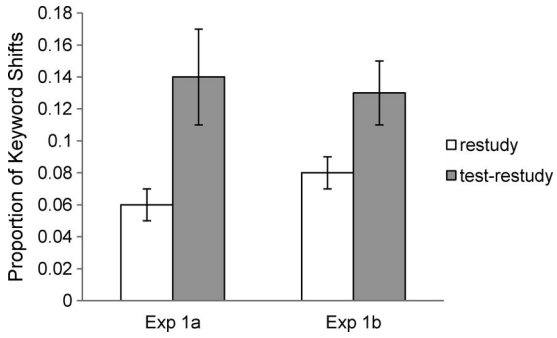

Figure 2. Mean proportion of keyword shifts during practice as a function of practice group in Experiments 1a and 1b. Error bars represent standard errors.

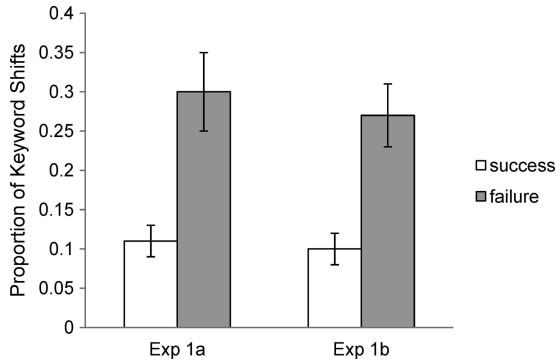

Figure 3. Mean proportion of keyword shifts during practice as a function of retrieval status during practice for the test–restudy groups in Experiments 1a and 1b. Error bars represent standard errors.

Experiment 2

The goals of Experiment 2 were threefold: Importantly, the first goal was to better evaluate Prediction 3 of the mediator shift hypothesis. The second goal was to increase the proportion of retrieval failures per item during practice by examining more difficult learning conditions. The third goal was to provide a stronger test of the mediator shift hypothesis by comparing two different test–restudy schedules of practice.

As in Experiments 1a and 1b, participants learned Swahili-English word pairs with restudy or test–restudy schedules of practice and returned 2 days later for a final cued-recall test and then a keyword recall test for all items. Importantly, the format of the keyword recall test in Experiment 2 was cue-only (as opposed to cue-target), which provides a better test of Prediction 3 because participants are required to retrieve keywords from memory. Additionally, the practice phase of Experiment 2 involved either a short lag (nine intervening items) or a long lag (69 intervening items) between each next practice trial with a given item. Based on previous research showing slower learning with longer lags (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review on the spacing effect literature), we expected a greater proportion of retrieval failures during practice for the long lag condition compared with the short lag condition in the test–restudy group. If so, the mediator shift hypothesis predicts a greater proportion of keyword shifts during practice for the long lag condition compared with the short lag condition in the test–restudy group.

Concerning effects of lag on keyword use in the restudy group, the most conservative prediction from the mediator shift hypothesis is that no differences will exist in the proportion of keyword shifts during practice. Because items are not being retrieved from

memory there would not be differences in retrieval failure during practice, which is presumably when keyword shifting occurs. However, a more liberal prediction (one that allows other mechanisms to influence metacognitive decisions during encoding) would allow for some effect of lag on keyword shifting in the restudy group.

Method

Participants and design. Participants included 55 Kent State University undergraduates who did not participate in Experiments 1a or 1b. Lag (short vs. long) was manipulated within participant. Practice group (test–restudy versus restudy) was manipulated between participants, with 29 and 26 participants randomly assigned to the test–restudy and restudy groups, respectively.

Materials. Items included 70 Swahili-English word pairs, with 60 target items and 10 filler items. The 60 target items included the 45 Swahili-English word pairs from Experiment 1b, with an additional 15 target Swahili-English word pairs, which were selected based on the pilot study described in Experiment 1. Items were randomly assigned to one of six practice lists. Three of these lists were assigned to the short lag condition, and the other three lists were assigned to the long lag condition, with list assignment to lag condition counterbalanced across participants. Ten fillers were used as primacy and recency buffer items.

Procedure. The general procedure was the same as Experiments 1a and 1b with the following exceptions. First, items received three restudy or test–restudy practice trials after their initial study and keyword report trials. Second, lag was a within participant manipulation. Third, on the final keyword recall test, only the cue was provided to prompt recall of the keyword(s) for each word pair.

Given that one goal of the current experiment was to increase the proportion of retrieval failures during practice, Experiment 2 was designed to optimize lag differences. Previous research has shown that longer versus shorter lags lead to more difficult learning and, therefore, will produce more retrieval failures during practice. Thus, with the goals of maximizing lag differences between conditions and keeping the length of the experiment manageable, the order of item presentation for the current experiment was blocked by list (see Appendix for detailed description of schedule of item presentation). Items in each of the three short lag lists were randomly sorted. The three lists assigned to the long lag condition were combined and randomly sorted. The five items assigned to be primacy items had an initial study trial, immediately followed by an initial keyword report trial. Primacy buffer items then had one practice trial (either restudy practice or test–restudy practice depending on the group to which the participant was assigned). After practice trials for the primacy buffer items, target items received an initial study trial, immediately followed by an initial keyword report trial. Initial study and keyword report trials were followed by three practice trials (either restudy or test–restudy practice), in which each restudy trial was immediately followed by a keyword report trial, as in Experiment 1. As illustrated in Appendix, the only difference between short lag and long lag items was that items assigned to the short lag lists completed the initial study trial and all practice trials (i.e., Practice Trials 1–3) for items in a block before moving on to the next list of to-be-learned items. For items in the long lag lists, each practice trial was followed by a block of trials for one set of short lag items. Upon completion of the last practice trial for long lag items, participants completed an initial study trial, initial keyword report trial, and one practice trial for the remaining five filler items to provide a recency buffer.

Results and Discussion

Mean final cued-recall performance as a function of lag condition and practice group is reported in Figure 4. Results of a 2 (Practice) × 2 (Lag) mixed-factor analysis of variance (ANOVA) showed a significant main effect of practice group, with higher levels of cued-recall for items learned with test–restudy practice versus restudy practice, $F(1, 53) = 32.76$, $MSE = 0.067$, $p < .001$. Results also showed a significant main effect of lag, with higher levels of cued recall for items learned with a long lag versus a short lag between practice trials, $F(1, 53) = 126.46$, $MSE = 0.013$, $p < .001$. The interaction was not significant ($F < 1$).

Before evaluating the predictions of the mediator shift hypothesis, a paired-samples t test was conducted to evaluate the extent to which the lag manipulation successfully increased the proportion of retrieval failures during practice for the test–restudy group. Results showed a significant difference between short lag and long lag conditions, with a greater proportion of retrieval failure trials during practice for items learned with a long lag versus a short lag (M proportion = .37, SE = .04, or about 1.1 out of 3 trials, vs. 0.20, SE = .04, or about 0.6 out of 3 trials), $t(28) = 8.29$, $p < .001$.

Keyword shifts during practice. The mean proportion of keyword shifts during practice was computed as in Experiments 1a and 1b and is reported in Figure 5. Supporting Prediction 1, results of a 2 (Practice) × 2 (Lag) mixed-factor ANOVA showed a significant main effect of practice group, with a greater proportion of keyword shifts during test–restudy practice versus restudy practice, $F(1, 53) = 8.32$, $MSE = 0.016$, $p = .006$. Results also showed a main effect of lag, with a greater proportion of keyword shifts for items learned with a long lag versus a short lag, $F(1, 53) = 19.50$, $MSE = 0.003$, $p < .001$. The interaction was not significant ($F < 1$).

The proportion of keyword shifts as a function of retrieval status during test–restudy practice is reported in Figure 6. Supporting Prediction 2, results of a 2 (Lag) × 2 (Retrieval Status) repeated-measures ANOVA showed a significant main effect of retrieval status, $F(1, 27) = 24.12$, $MSE = 0.017$, $p < .001$. Keywords were

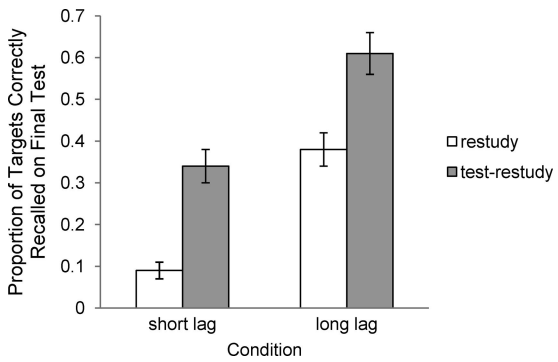

Figure 4. Mean proportion of targets correctly recalled on the final cued recall test as a function of practice group and lag condition in Experiment 2. Error bars represent standard errors.

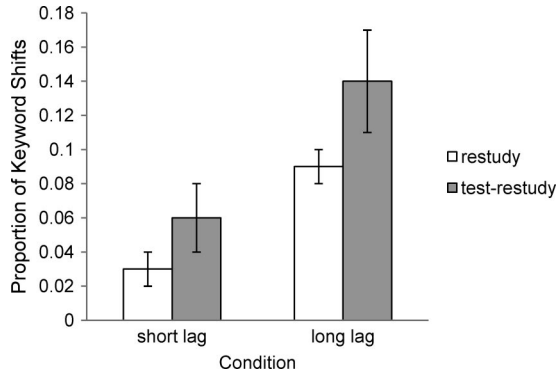

Figure 5. Mean proportion of keyword shifts during practice as a function of practice group and lag condition in Experiment 2. Error bars represent standard errors.

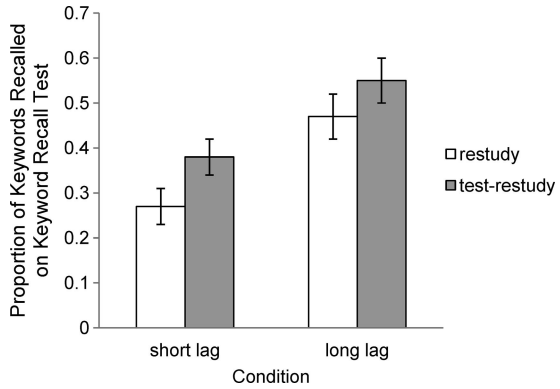

Figure 7. Mean proportion of keywords recalled on the keyword recall test as a function of practice group and lag condition in Experiment 2. Error bars represent standard errors.

more likely to shift after retrieval failure trials versus retrieval success trials. The main effect of lag approached significance, $F(1, 27) = 3.52$, $MSE = 0.013$, $p = .07$. The interaction was not significant ($F < 1.3$).

Keyword recall test. To revisit, one important goal of the current experiment was to provide a better test of Prediction 3 by implementing a cue-only prompt for the keyword recall test. The mean proportion of keywords recalled on the keyword recall test as a function of lag condition and practice group is reported in Figure 7. Consistent with Prediction 3, a 2 (Practice) × 2 (Lag) mixed-factor ANOVA showed a significant main effect of practice group, with a greater proportion of keywords recalled on the keyword recall test for items learned with test–restudy practice versus restudy practice, $F(1, 53) = 12.41$, $MSE = 0.075$, $p = .001$. Results also showed a main effect of lag, with a greater proportion of keywords recalled for items learned with a long lag versus a short lag, $F(1, 53) = 89.48$, $MSE = 0.003$, $p < .001$. The interaction was not significant, $F < 1.7$. These results confirm that test–restudy practice leads to the generation of more recallable mediators, one of the two key dimensions of mediator effectiveness.

Additionally, the advantage in keyword recall for test–restudy versus restudy was due almost entirely to greater recall of keywords used at the end of practice (which presumably include relatively effective mediators), rather than to greater recall of less effective keywords that were used earlier in practice and subsequently discarded. Indeed, in both groups, almost all of the keywords recalled were from the end of practice. Collapsing across lag, the mean number of practice-final keywords recalled was 31.3 in the test–restudy group and 22.3 in the restudy group, $t(53) = 1.92$, $p = .06$. In contrast, mean number of nonfinal keywords recalled was 1.9 in the test–restudy group and 0.7 in the restudy group, $t(53) = 2.67$, $p = .01$.

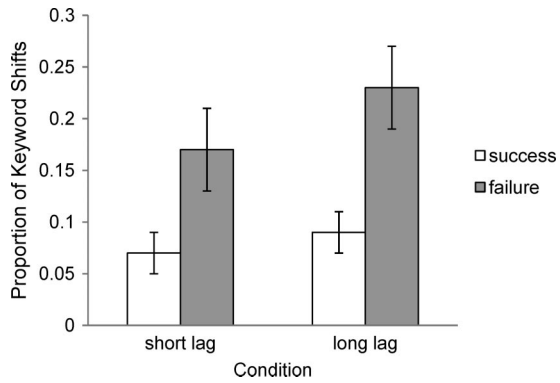

Figure 6. Mean proportion of keyword shifts during practice as a function of retrieval status during practice for conditions of the test–restudy group in Experiment 2. Error bars represent standard errors.

To provide evidence that the keywords used later in practice were in fact more effective, we conducted post hoc conditional analyses of target word recall on the final cued recall test as a function of the keyword recalled for that item on the keyword recall test. For each subset of keywords (keywords used on the final practice trial versus nonfinal keywords), we computed the proportion of targets that were correctly recalled on the final cued recall test. If keywords used on the final trial during practice (vs. earlier trials) are more effective, they will be more likely to elicit retrieval of the target. (Given the low level of nonfinal keyword recall, we collapsed across lag for these analyses to reduce the number of participants that would otherwise be excluded because of missing values.) Consistent with this expectation, final cued recall performance was substantially greater for items with final keywords recalled versus nonfinal keywords recalled (60% vs. 5%), $F(1, 18) = 61.09$, $MSE = 0.024$, $p < .001$. The main effect of practice group did not reach significance ($F < 2.4$), nor did the interaction ($F < 1.6$).

In sum, results replicated the pattern of results from Experiments 1a and 1b, supporting Prediction 1 and Prediction 2 of the mediator shift hypothesis. Importantly, when a more appropriate cue-only keyword recall test was utilized, results also supported Prediction 3.

In addition to supporting predictions of the mediator shift hypothesis, the lag manipulation provided a further test of the hypothesis by comparing two test–restudy conditions, which were predicted to differentially influence final test performance and keyword shifting during practice. Replicating previous spacing effect research, a greater proportion of items were correctly recalled on the final cued-recall test for items learned with a long lag versus a short lag. Importantly, the pattern of results for keyword

shifting during practice and for keyword recall on the final keyword recall test were in line with predictions from the mediator shift hypothesis. Specifically, a greater proportion of keyword shifts occurred during practice for items learned with a long lag versus a short lag. Similarly, a greater proportion of keywords from Session 1 were recalled on the final keyword recall test for items learned with a long lag versus a short lag. These results are consistent with higher levels of final test performance for the long lag condition compared with the short lag condition in the test–restudy group.

Similar effects of lag on keyword shifting during practice and keyword recall on the final keyword recall test obtained in the restudy group, which is not predicted by the mediator shift hypothesis. These results are perhaps not surprising, given that retrieval failure is likely not the only factor underlying shifts to more effective mediators. Importantly, however, although the qualitative pattern of results was similar for conditions of the restudy and test–restudy groups, keyword shifting and keyword recall was greater for items learned with test–restudy practice versus restudy practice. These results are consistent with the idea that, above and beyond any other shared factors, retrieval failures during test–restudy practice further facilitate shifting to more effective mediators.

General Discussion

Across a series of three experiments, we evaluated predictions of the mediator shift hypothesis for explaining the memorial benefits of test–restudy versus restudy practice. According to the mediator shift hypothesis, retention is greater after test–restudy practice versus restudy practice because individuals modify mediators when they cannot correctly recall an item during practice. Individuals engaged in restudy practice cannot evaluate the effectiveness of mediators as well as individuals engaged in test–restudy practice because they do not experience retrieval failure during encoding. Across experiments, three predictions of the mediator shift hypothesis were confirmed. Supporting Prediction 1, a greater proportion of keyword shifts occurred during test–restudy practice versus restudy practice. Supporting Prediction 2, a greater proportion of keyword shifts occurred after retrieval failure trials versus retrieval success trials during test–restudy practice. Supporting Prediction 3, the proportion of keywords from Session 1 recalled on the keyword recall test in Session 2 was greater for items learned with test–restudy practice versus restudy practice in Experiment 2.

Thus, the current experiments demonstrated that keywords were more likely to shift during practice for items learned with test–restudy practice versus restudy practice, presumably because retrieval failures during practice allowed individuals to evaluate the effectiveness of keywords and to shift to more effective ones. These more effective keywords facilitated recall of targets on the final retention test. To revisit, previous research has demonstrated that the effectiveness of mediators depends on at least two factors, mediator retrieval and mediator decoding (e.g., Dunlosky et al., 2005). For a mediator to be effective it must be retrieved when prompted by a cue on a test trial. Additionally, the mediator must be decoded to elicit the correct target response.

Pyc and Rawson (2010) recently provided evidence for a contribution of both of these factors to the memorial benefits of testing. In brief, they tested the *mediator effectiveness hypothesis*, which states that mediators generated during test–restudy practice (compared with restudy practice) are more likely to be retrieved and decoded, which facilitates recall of target answers on a retention test for items learned with test–restudy practice versus restudy practice. Pyc and Rawson had participants learn Swahili-English word pairs using the keyword encoding strategy and a procedure similar to the one used in the current experiments, but they used different final test formats to examine both mediator retrieval and mediator decoding. Concerning mediator retrieval, participants in one group were prompted with the cue word on each final test trial and were asked to retrieve their own mediator (i.e., keyword) for a given item prior to recalling the target answer. Mediator recall was greater for items learned with test–restudy practice versus restudy practice. Concerning mediator decoding, participants in another group were prompted with the cue word on each final test trial but were also provided with the last keyword that they had generated for a given item in the practice session. Recall of targets was greater for items learned with test–restudy practice versus restudy practice. Thus, when differences in mediator recall were eliminated (because participants were provided with their own mediators), mediators generated in the test–restudy group (vs. the restudy group) were more likely to elicit the correct target answer.

The mediator shift hypothesis complements the mediator effectiveness hypothesis by stating that more effective mediators arise at least in part from shifting mediators after retrieval failure trials during practice. Taken together, these two hypotheses suggest that mediator use during test practice is one theoretical explanation for the memorial benefits of test–restudy versus restudy practice. That is, shifting mediators after retrieval failure trials during practice leads to the generation of more effective mediators (i.e., mediators that are more likely to be retrieved and decoded), which subsequently facilitates retrieval of target answers.

The current experiments provided support for predictions of the mediator shift hypothesis and may actually provide an underestimate of the potential role of mediator shifting. The particular way that participants were instructed to use the keyword encoding strategy could have limited the extent to which keywords could shift during practice. Participants were instructed to generate an English word (i.e., keyword mediator) that sounded like the Swahili cue and then to semantically relate the keyword to the English target. Constraining participants to semantically relate keywords to English targets may have limited the number of keyword mediators participants could generate for a given item during encoding because it may have been difficult to generate keywords that sound like the Swahili word and that are semantically related to the English target. By contrast, instructing participants to generate an English keyword but then form an image including the English keyword and the English target may have afforded a greater number of viable mediators (and hence more shifting) during practice. Almost any two English words can be formed into an image, whereas it is much more difficult to semantically relate any two English words. Thus, future research evaluating the mediator shift hypothesis could use a variation of the keyword strategy used here (e.g., one in which individuals form an image including the keyword and the English target), which may lead to even stronger support for predictions of the hypothesis.

The current studies may also underestimate of the role of mediator shifting in that we limited learners to the use of only one kind of encoding strategy (i.e., the keyword method). Although

this strategy is effective for learning foreign language vocabulary, several other effective encoding strategies would also permit shifting from less effective to more effective mediators. For example, imagery is a useful strategy for learning various kinds of material (e.g., word lists, paired associates, concepts, text), but the effectiveness of imagery depends on the nature of the image formed. For the word pair *dog–spoon*, a noninteractive image of a dog sitting next to a spoon would be less effective than an interactive image of a dog eating with a spoon. Likewise, for a student attempting to learn from a text describing a complex process or mechanism (e.g., bat echolocation, how the braking system of a car works), a static image of the components or parts involved presumably would be less effective than a dynamic image of the parts engaged in the process. Of course, other encoding strategies may be employed for different kinds of materials (e.g., compare-contrast or example generation are useful strategies for learning the kinds of key terms or concepts encountered in many course materials), but the important point is that most of these strategies afford more versus less effective instantiations of the kind of mediator involved in that strategy (e.g., comparing-contrasting along relevant vs. irrelevant dimensions). Thus, an important direction for future research will be to evaluate the extent to which the effects observed here generalize to other encoding strategies and to other kinds of materials and tasks.

Researchers have proposed two effects of testing (i.e., direct and mediated), which may be influenced by different underlying mechanisms. To revisit, the direct effects of testing involve benefits from the act of testing itself. The mediated effects of testing involve the influence of testing on subsequent encoding during restudy. The mediator shift hypothesis provides one theoretical explanation for the mediated effects of testing, in that the outcome of the practice tests presumably improved memory by influencing the generation of mediators during the subsequent restudy opportunity. However, it is important to note that the design of the current studies precludes us from concluding that we only observed mediated effects of testing. It is likely that the difference in final cued recall for test–restudy versus restudy was also due in part to direct effects of testing. This possibility suggests that another fruitful direction for future research will be to explore how mediator shifting may work in concert with other theoretical mechanisms that have been proposed (such as elaborative retrieval and retrieval effort, as described in the Introduction).

To conclude, the current experiments evaluated and confirmed predictions of the mediator shift hypothesis. These results suggest that one reason testing is beneficial for memory is because testing during practice affords an evaluation of the effectiveness of mediators. On retrieval failure trials during practice, individuals evaluate the effectiveness of mediators and shift from less effective mediators to more effective mediators on subsequent restudy trials, which facilitates retrieval of information on a later retention test. Most important, the current studies support one theoretical explanation for testing effects, which are empirically well established but theoretically not well understood.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. doi:10.1002/acp.1391

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566–577. doi:10.1016/j.jml.2005.01.012

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.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941–956. doi:10.1002/acp.1239

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 1563–1569. doi:10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276. doi:10.3758/BF03193405

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14,* 474–478. doi:10.3758/BF03194092

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36,* 438–448. doi:10.3758/MC.36.2.438

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. doi:10.1037/0033-2909.132.3.354

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology, 41,* 389–400. doi:10.1037/0012-1649.41.2.389

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign language vocabulary learning. *Applied Cognitive Psychology, 21,* 499–526. doi:10.1002/acp.1287

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology, 60,* 991–1004. doi:10.1080/17470210600823595

Karpicke, J. D., & Roediger, H. L., III (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

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

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15,* 257–280. doi:10.1080/13825580701322171

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representation through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. doi:10.1037/0278-7393.11.2.371

McDaniel, M. A., Pressley, M., & Dunay, P. K. (1987). Long-term retention of vocabulary after keyword and context learning. *Journal of Educational Psychology, 79,* 87–89. doi:10.1037/0022-0663.79.1.87

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2,* 325–335. doi:10.1080/09658219408258951

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35,* 1917–1927. doi:10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Raugh, R. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology, 67,* 1–16. doi:10.1037/h0078665

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Roediger, H. L., III., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Appendix

Table A1
Schedule of Item Presentation in Experiment 2

No. of items	Block of practice																									
10	PB_1	PB_2																							RB_1	RB_2
30			A_1					A_2							A_3								A4			
10				B_1	B_2	B_3	B_4																			
10									C_1	C_2	C_3	C_4														
10																D_1	D_2	D_3	D_4							

Note. PB = primacy buffer items; RB = recency buffer items. Number in subscript refers to trial number, in which 1 = initial study and initial keyword trials and 2–4 = subsequent practice trials (either restudy or test–restudy). Letters A–D refer to set of items.

Received September 17, 2010
Revision received September 1, 2011
Accepted October 1, 2011 ■