

Examining the efficiency of schedules of distributed retrieval practice

MARY A. PYC AND KATHERINE A. RAWSON
Kent State University, Kent, Ohio

Given that students typically have a sizeable amount of course material to learn but a finite amount of study time, evaluating the efficiency of study schedules is important. We explored the efficiency of various schedules of distributed retrieval plus restudy. Across two experiments, 227 undergraduates were asked to learn Swahili–English vocabulary word pairs. In *conventional* schedule groups, all items were presented for 3 practice trials after initial study (as in most previous research). In *dropout* schedule groups, the number of practice trials allocated to each item varied, in that practice with a given item was discontinued after criterion performance had been reached. A dropout schedule led to levels of performance similar to those for conventional schedules (but in fewer trials), and it was particularly effective for learning initially incorrect items. However, the efficiency of the various schedules depended critically on the interval between presentations of an item. Results suggest that dropout can be a more efficient learning schedule for students than can conventional schedules of practice.

From the earliest years of schooling, students are faced with sizeable amounts of course material that they are expected to master. However, students have a finite amount of time and energy to invest in learning the required content. Thus, they must efficiently manage their workload if they are to master all course materials. Being able to successfully manage workload is especially important as students progress through school, where they are responsible for learning increasing amounts of information across various disciplines. Given the amount of material that students must learn and the limited amount of time they have to invest, an important question arises: How efficient are the study strategies that are recommended for improving students' learning? The highest level goal of the present research was to explore how to promote student learning of course material by the most efficient means possible. At the most general level, in the present article, *efficiency* refers to spending the least amount of time necessary to learn a given item well enough to be able to remember that item later.

Effective student learning may be described in terms of the amount of information learned (memory level) or in terms of the efficiency with which it is learned (time and effort expended). Of these two, the primary focus of previous research has overwhelmingly been on overall memory level (see, e.g., Cull, 2000; Morris, Fritz, Jackson, Nichol, & Roberts, 2005; Pashler, Zarow, & Triplett, 2003; Roediger & Karpicke, 2006a), with an emphasis on investigation of study conditions that improve memory for target information. Among the various study conditions that have been examined in previous research, two conditions have repeatedly been shown to have robust effects on memory: (1) distributed practice and (2) retrieval practice followed

by restudy. First, distributed practice involves multiple presentations of an item with intervening time and material between each, as opposed to massed practice in which all presentations occur immediately following one another. The advantage of distributed practice over massed practice (i.e., the spacing effect) is one of the most well-established findings in experimental psychology, with distributed practice yielding better memory than massed practice with different kinds of materials, tasks, and test delays (for recent reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Roediger & Karpicke, 2006b).

Second, an increasing amount of research has shown that practice retrieving target items from memory followed by restudy of those items leads to higher levels of memory than either activity alone, even when controlling for the amount of time spent on a task (see, e.g., Carpenter & DeLosh, 2005; Carrier & Pashler, 1992; Cull, 2000; Cull, Shaughnessy, & Zechmeister, 1996; Morris et al., 2005; Roediger & Karpicke, 2006a; Wheeler, Ewers, & Buonanno, 2003).

Most important, recent research has shown that these two conditions together—that is, retrieval practice plus restudy implemented on a distributed schedule—is a potent combination for improving memory (Bahrnick & Hall, 2005; Carpenter, & DeLosh, 2005; Cull, 2000; Cull et al., 1996; Landauer & Bjork, 1978; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler et al., 2003; Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005; Wheeler et al., 2003). For example, Cull (2000) presented paired associates for an initial study trial, followed by three practice trials that were either massed or distributed. Within each schedule condition, practice trials involved restudying the word

pairs (restudy only), the presentation of the cue word as a prompt to retrieve the target word (retrieval only), or retrieval followed by restudy.¹ A short filler task was followed by a final cued-recall test. Results showed that distributed practice yielded higher final test performance than did massed practice. Additionally, retrieval plus restudy yielded higher final test performance than did either retrieval or restudy alone. Most important, the distributed retrieval plus restudy condition yielded the highest level of performance (49%, in comparison with an average of 26% across all other groups)—a finding replicated in follow-up experiments involving different materials and different distributed practice intervals (from minutes to days)—with performance as high as 98% in one distributed retrieval plus restudy condition (in comparison with 60% on average across other conditions).

Thus, previous research has firmly established that distributed retrieval plus restudy can be highly effective for improving overall performance level. In contrast, the efficiency of distributed retrieval plus restudy schedules has not been systematically investigated; hence, it is less well understood. Specifically, what is the most *efficient* schedule of distributed retrieval plus restudy practice? Given the amount of information that students must learn, it is important for one to consider both the absolute level of performance obtained and the efficiency with which it is obtained when evaluating the effectiveness of a practice schedule. Accordingly, the goal of the present research was to investigate the efficiency of various schedules of distributed retrieval plus restudy.

A given practice schedule can be more efficient than another in one of two ways. First, a practice schedule may yield a higher level of memory for target information while using the same number of practice trials as another schedule. Second, a practice schedule may yield a similar level of memory while using fewer practice trials than another schedule. Although seldom couched in terms of efficiency, the results of previous research bear on the first conceptualization of efficiency. In most previous research, items in different practice conditions have received the same number of trials, with the main manipulation involving the timing and/or activity performed during those trials. As summarized previously, a distributed schedule typically yields a higher level of memory than does a massed schedule, using the same number of trials. Similarly, a schedule of retrieval plus restudy yields a higher level of memory than either activity alone, using the same number of trials (and controlling the total amount of time on task during practice). In this sense, the efficiency of distributed retrieval plus restudy schedules has been partly established (for ease of exposition, we hereafter refer to schedules of distributed retrieval plus restudy practice simply as *test–restudy schedules*).

Importantly, however, the test–restudy schedules that were implemented in previous research have assigned the same number of test–restudy trials to each item and each learner (hereafter referred to as *conventional* schedules). To what extent might the efficiency of test–restudy schedules be further improved by allowing variability in the number of trials allocated to particular items? That is, a schedule

in which fewer trials are allocated to recallable items and more trials are allocated to unrecallable items may be more efficient, in that higher levels of memory may be achieved using the same number of trials. The idea of using a variable schedule is also relevant to the second conceptualization of efficiency, in that a variable test–restudy schedule may yield a similar level of memory while using fewer trials than a conventional schedule. Given that previous research has not directly compared variable schedules with conventional schedules, the extent to which students can achieve a similar level of learning (or better) with fewer test–restudy trials is currently unknown.

The paucity of research on variable test–restudy schedules is somewhat ironic, given that this strategy is likely one that is spontaneously used by many students. That is, a strategy often used by students to learn course content is to use “flashcards,” by which they attempt to recall some information from a prompt on one side of the card (test) and then check the correct answer on the opposite side of the card (restudy). Although we were unable to find research that systematically evaluates the schedules that students typically use when studying with flashcards, an informal survey of our own students—as well as our intuition—suggest that students typically adopt a variable schedule of practice. In other words, students are more likely to remove cards from their stack as they learn items to spend more time on those items that are not yet learned, rather than cycling through every card in the stack a set number of times. Related findings come from research evaluating models of self-regulated learning, showing that the number of trials and/or amount of time individuals allocate to particular items varies with the perceived difficulty of learning those items (see, e.g., Thiede & Dunlosky, 1999).

An important question follows from these observations: Are variable test–restudy schedules more efficient than conventional schedules? That is, do variable schedules yield either higher levels of memory in the same number of trials or a similar level of memory in fewer trials? The idea of borrowing some trials from recallable items to spend on unrecallable items in a variable schedule is intuitive. However, research on overlearning (i.e., “immediate continuation of practice beyond criterion of one perfect instance,” Rohrer et al., 2005, p. 361) suggests that there may be some cost to discontinuing practice of initially recallable items (see, e.g., Driskell, Willis, & Copper, 1992; Krueger, 1929; Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982; Rohrer et al., 2005). For example, Nelson et al. (1982) found that final test performance was greater for paired associates that were practiced to four correct cued recall responses than was performance for items that were practiced to two correct responses, which in turn was greater than performance for items that were practiced to one correct response. Thus, one possibility is that any gains in learning from spending additional trials on unrecallable items may be offset by losses from discontinuing practice of recallable items. If so, then the additional practice trials allocated to recallable items in conventional schedules may lead to overall higher levels of memory in comparison with memory in a variable schedule in which practice with recallable items is discontinued.

Given the importance of these issues and the lack of a clear answer from previous research, we compared the efficiency of a *dropout* method for scheduling test–restudy trials to conventional schedules that have been used in previous research. Specifically, for the dropout method, items that were correctly recalled during a test–restudy trial were removed from the list of to-be-practiced materials. As a result, the trials that would normally have been used for further practice with those items (in a conventional schedule) could be allocated to items that an individual had more difficulty mastering. Although dropout schedules have been used to measure trials to criterion in savings of learning studies (see, e.g., Bahrck, 1967; Krueger, 1929), they have been neither competitively evaluated against conventional test–restudy schedules nor systematically explored as a means for promoting efficient learning more generally.

Accordingly, we compared dropout and conventional test–restudy schedules in two experiments. To foreshadow, results from Experiment 1 showed that a dropout schedule can be more efficient than conventional schedules, yielding equivalent levels of memory in fewer trials. Experiment 2 was conducted to replicate and extend the results from Experiment 1.

EXPERIMENT 1

In Experiment 1, all participants were asked to learn Swahili–English paired associates. Items were first presented for an initial study trial and then underwent one of five test–restudy schedules, with two conventional schedules and three dropout schedules. In the conventional schedule groups, each item had three test–restudy trials with an average of five intervening items between presentations of a given item, which is an interstimulus interval (ISI) that has commonly been used in previous research with conventional schedules (see, e.g., Cull, 2000).

In the dropout groups, the number of test–restudy trials per item was variable. In the *5-drop* group, we used an ISI of five items between initial study and the first test–restudy trial for each item, comparable to the ISI used in the conventional schedules. If an item was correctly recalled on the first trial, the item was dropped from further practice. If an item was not correctly recalled, it was repeated at the end of the list of to-be-learned items until one correct recall was obtained. The schedule in the *5-drop-after-2* group was motivated by the overlearning literature suggesting that recallable items benefit from additional practice. In this group, the initial ISI was also five items as in the *5-drop* and conventional schedule groups, but an item was not dropped from the list of to-be-learned materials until it had been correctly recalled twice. Finally, in the *23-drop* group, the initial ISI for each item was 23. We implemented an ISI of 23 based on research by Pashler et al. (2003) indicating that even though items practiced after a longer ISI may be incorrectly recalled more often at first test, they have an advantage over shorter ISIs for later tests. As in the *5-drop* group, once an item was correctly recalled on one practice trial, it was dropped from the list of to-be-learned items. Measures of primary interest were performance on a final cued recall test and the number of

test–restudy trials used during practice to obtain that level of performance.

Method

Participants and Design. A total of 161 undergraduate students enrolled in Introductory Psychology at Kent State University participated in return for course credit. Participants were randomly assigned to one of five groups: 5–5–5, 1–5–9, 5-drop, 5-drop-after-2, and 23-drop, with 31–33 participants in each group. Test–restudy or study only (described further below) was a within-subjects variable.

Materials. Two lists of 24 Swahili–English translation word pairs were used, with an equivalent range of item difficulty on each list (based on norms reported by Nelson & Dunlosky, 1994).

Procedure. All task instructions and items were presented via computer. All items in all groups received an initial study trial. On the initial study trial, the Swahili word was presented on the left of the screen and the target English translation appeared on the right for 10 sec. For each participant, 24 word pairs received test–restudy practice after initial study. During a test–restudy trial, the Swahili word was presented alone, and participants had 8 sec to enter the English translation in a text box provided below the Swahili word. After 8 sec, the response box was removed from the screen, and the Swahili and English words were presented together for 4 sec.

We included two conventional test–restudy schedules.² In the 5–5–5 group, each item was presented for three test–restudy trials, with an ISI of five items between initial study and each test–restudy trial. In the 1–5–9 group, each item was presented for three test–restudy trials, with an ISI of one, five, and nine items between initial study and each test–restudy trial.

In the *5-drop* group, five items intervened between initial study and the first test–restudy trial. If the correct translation was recalled, the item was dropped from further practice. If the item was not correctly recalled, then it was placed at the end of the list of items for another test–restudy trial. This process continued until (1) all items were correctly recalled once or (2) a participant reached the 72 maximum-trial allowance. As described previously, we were interested in whether dropout schedules could yield either higher levels of memory in the same number of trials or a similar level of memory in fewer trials than the conventional schedules. Thus, we set the maximum allowance of test–restudy trials in the dropout groups at 72 in order to match the 72 test–restudy trials allotted in the conventional schedules. This stopping rule affected 13 participants in the *5-drop* group, 30 participants in the *5-drop-after-2* group, and 17 participants in the *23-drop* group. Note that the initial study trials did not count toward the maximum of 72 total test–restudy trials.

In the *5-drop-after-2* group, the initial ISI was five items, as in the *5-drop* group. However, this group differed in that an item had to be correctly recalled two times before it was dropped. After the first test–restudy trial, all items were placed at the end of the list of to-be-learned items. If an item was correctly recalled on the first two test–restudy trials, it was dropped from the list; otherwise, it was again placed at the end of the to-be-learned list of items. This process continued until either (1) an item had been correctly recalled on two test–restudy trials or (2) the participant reached the 72 test–restudy trial maximum allowance. The two correct recalls needed for an item to be dropped did not have to occur on consecutive test–restudy trials.

The procedure in the *23-drop* group was identical to that in the *5-drop* group, with the exception that the initial ISI was 23 items. We should note that although the ISI between initial study and the first test–restudy trial in each dropout group was fixed, the number of intervening items between subsequent test–restudy trials was variable, because items that needed additional practice were placed at the end of the list of the to-be-learned items. Therefore, in the *5-drop-after-2* and the *5-drop* groups, although the ISI between initial study and the first test–restudy trial was 5, the lag between subsequent test–restudy trials was greater than five (across participants, the mean ISI between test–restudy trials was 28, 11, 10, 6, and 3 items in the *5-drop* group, and 24, 20, 16, and 2 intervening items in the *5-drop-*

after-2 group). In contrast, for the 23-drop group, the lag between subsequent test–restudy trials decreased from 23 to 21, 18, 13, and 3 intervening items.

The other 24 word pairs were used as study-only controls, where items received no further practice after initial study. Implementing the 1–5–9 schedule necessitated the use of filler items in order to maintain an expanding interval between trials for all items. Given that this schedule included filler (control) items, we also presented them in the other groups in order to equate the total number of items presented. Because the schedules in the other groups would not allow embedding of the control items, 12 control items were presented before the test–restudy items, and 12 were presented after the test–restudy items. Appendix A contains more specific details about the item presentation schedule for each group. Assignment of the two 24-item lists to condition was counterbalanced across participants.

For all groups, the practice phase was followed by an experimenter-paced reading-comprehension filler task (unrelated to the materials being studied and tested) that took 40 min to complete. A final cued-recall test was then administered. Final cued recall was self-paced.

Results and Discussion

To revisit, our primary goal was to compare the efficiency of dropout schedules with conventional schedules of test–restudy practice. Accordingly, statistical tests were focused on the comparison of each of the dropout schedules with each conventional schedule in order to examine the extent to which dropout schedules yielded either higher levels of memory in the same number of trials or a similar level of memory in fewer trials.

As a reminder, both the 1–5–9 and 5–5–5 groups were given a set number of trials (72), with each of the 24 items receiving three test–restudy trials. In contrast, the number of test–restudy trials used in the dropout groups was variable. The mean number of test–restudy trials used across participants in each group is reported in the top panel of Figure 1. Set values for the 1–5–9 and 5–5–5 groups are included in the figure for ease of comparison with the dropout groups. Results of one-sample *t* tests indicated that the 5-drop, 23-drop, and 5-drop-after-2 groups used significantly fewer than 72 trials [$t(31) = 4.99, p < .001$; $t(32) = 3.38, p = .002$, and $t(32) = 1.76, p = .044$, respectively]. Thus, all three dropout schedules used fewer trials than did the conventional schedules of practice, although the difference in the 5-drop-after-2 group is arguably trivial.

Of course, using fewer trials to learn items during a practice session is preferable only if there is not a substantial loss of that information over the retention interval. The mean percentage of items correctly recalled at final test for each group is reported in the bottom panel of Figure 1. For ease of comparison of the qualitative patterns across groups in the two key measures (number of trials used for test–restudy items and final test performance for these items), performance for the test–restudy items is displayed in the bars of the figure; performance for study-only control items is reported in parentheses within the bar for each group. A 2 (activity: test–restudy vs. study-only) \times 5 (schedule) mixed-factor ANOVA indicated a main effect of activity [$F(1,156) = 767.03, p < .001$]. The finding that test–restudy items had significantly higher performance than did study-only items is not surprising and replicates previous research that has used study-only control conditions (see, e.g., Carpenter, Pashler, & Vul,

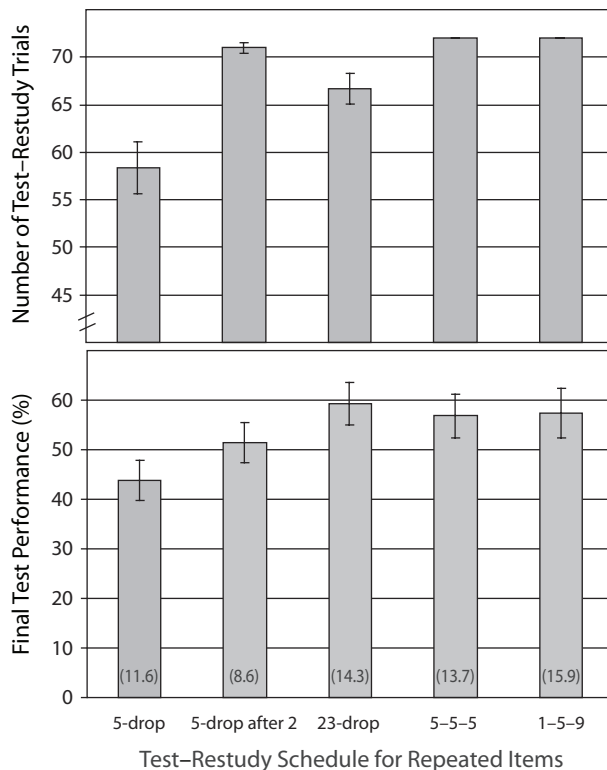


Figure 1. Top panel: Mean number of test–restudy trials used for repeated items across participants in each group in Experiment 1. Error bars represent standard errors. Number of trials was set at 72 in the 5–5–5 and 1–5–9 groups; these set values are presented for purposes of comparison with the dropout groups. Bottom panel: Mean percentage of repeated items correctly recalled on final test across participants in each group in Experiment 1. Error bars represent standard errors. Values in parentheses represent percentage of control items correctly recalled on final test.

2006; Cull, 2000). The main effect of schedule was not significant [$F(4,156) = 1.76, p = .14$]. The interaction approached significance [$F(4,156) = 2.33, p = .058$]. Follow-up tests showed that performance in the 5-drop group was lower than that in the 5–5–5 and 1–5–9 groups [$t(62) = 2.19$ and $t(61) = 2.10$, respectively], a finding we will consider further in the General Discussion. Performance in the other two dropout groups did not differ significantly from that in either of the two conventional groups ($ts < .91$), with the 23-drop group showing a slight numerical advantage over the conventional groups.

Examination of the overall pattern in Figure 1 suggests that the 23-drop group was the most efficient, given that significantly fewer test–restudy trials were used to achieve the same level of performance as that in the conventional schedule groups. One could further combine these two raw measures (trials used and final performance) into a derived efficiency measure by dividing the percentage of items recalled on the final test by the number of test–restudy trials used during practice for each participant. Mean recall per trial (and standard error) across participants in each group was .79 (.06), .80 (.07), .93 (.14), .94 (.09), and .74 (.06) in the 5–5–5, 1–5–9, 5-drop, 23-drop,

and 5-drop-after-2 groups, respectively. Note, however, that the derived measure of efficiency would suggest that the 5-drop and 23-groups are equally preferable to the conventional schedules, which is clearly not the case when one considers the absolute level of performance achieved in each group. Thus, for present purposes, we believe that an examination of the raw measures is most informative.

To further compare the dropout and conventional schedules, we examined the extent to which these schedules differentially benefited initially recallable and unrecallable items. Specifically, we conducted conditional analyses to examine final test performance as a function of the kind of response generated for an item on the first test–restudy trial (either correct or incorrect). Results for each group are reported in Table 1.³ As with the measures above, we compared each of the dropout schedules with each conventional schedule. Results indicated that in the 5-drop group, initially correct items were significantly less likely to be correct at final test than they were in the 5–5–5 and 1–5–9 groups [$t(61) = 6.21, p < .001$ and $t(61) = 4.72, p < .001$, respectively]. In contrast, final test performance for initially correct responses for the other two dropout groups did not differ significantly from either conventional schedule, ($ts < 1.2$). Thus, for the 23-drop and 5-drop-after-2 schedules, items that were correctly recalled on the first test–restudy trial were just as likely to be correct at final test as with conventional schedules, despite not having been given the additional test–restudy trial(s) that were administered in the conventional schedules.

Presumably, in the dropout groups, some of the test–restudy trials that were made available by dropping initially correct items from further practice could be used for additional practice with initially incorrect items. Thus, one might expect that final test performance for initially incorrect items would be greater in the dropout groups than in the conventional schedule groups. Indeed, this was the case for the 23-drop group, with initially incorrect items more likely to be correct at final test than in the 5–5–5 and 1–5–9 groups [$t(63) = 1.90, p = .067$ and $t(61) = 2.02, p = .048$, respectively]. A straightforward interpretation of this advantage is that some of the initially incorrect items received additional practice in the 23-drop group. Consistent with this idea, on average, across participants in the 23-drop group, 5.1 of the initially incorrect items received

four or more test–restudy trials (with some receiving up to six). Final test performance for initially incorrect responses for the other two dropout groups did not differ significantly from either conventional schedule ($ts < 1.4$).

Note that for these conditional analyses, we have reported performance on the first test–restudy trial during practice. For archival purposes, we report performance on all test–restudy trials in Appendix B.

Taken together, the results of Experiment 1 show that not all variable schedules of test–restudy practice will be more efficient than conventional schedules. Although the 5-drop schedule used significantly fewer trials than did the conventional schedules, it yielded lower levels of memory. The 5-drop-after-2 schedule approached the same number of trials as those in the conventional schedules, but did not yield a higher level of memory. Thus, neither of these schedules is to be preferred to the conventional schedules examined here.

In contrast, the 23-drop schedule was a strong competitor for the conventional schedules of learning, in that it yielded a similar level of memory in significantly fewer trials. The conditional analyses further indicated that the 23-drop schedule was particularly effective for learning initially incorrect items. Presumably, the advantage of the 23-drop schedule for initially incorrect items was due to the allocation of additional test–restudy trials to those items. However, we address alternative interpretations of this result in Experiment 2.

EXPERIMENT 2

Experiment 1 provided initial evidence that a variable test–restudy schedule can be more efficient than conventional schedules, yielding similar memory levels in fewer trials. The 23-drop schedule was particularly advantageous for items that were incorrect on the first test–restudy trial, leading to the highest levels of performance on the final test for these items. A straightforward interpretation of this finding is that the 23-drop schedule benefited initially incorrect items via allocation of the additional trials made available by dropping initially correct items from further practice.

However, the advantage of the 23-drop schedule for initially incorrect items could reflect item difficulty dif-

Table 1
Final Test Performance As a Function of Kind of Response on
First Test–Restudy Trial for Repeated Items, Experiment 1

Group	Correct Recall on First Test–Restudy Trial				Incorrect Recall on First Test–Restudy Trial			
	No. Items Correctly Recalled		% Correct Recall on Final Test		No. Items Incorrectly Recalled		% Correct Recall on Final Test	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
5-drop	11.0	0.83	31.9	4.5	13.0	0.83	54.6	5.2
5-drop-after-2	10.8	0.81	60.3	4.9	13.2	0.81	44.1	4.5
23-drop	3.4	0.60	71.3	7.0	20.6	0.60	58.4	4.4
5–5–5	14.2	0.86	67.6	3.6	9.8	0.86	45.9	5.1
1–5–9	16.5	0.84	63.2	4.9	7.5	0.84	42.8	6.5

Note—Total number of items in each group was 24.

ferences, since the groups differed in the number of items that were incorrect on the first test–retest trial. Consider performance on the first test–retest trial in the 23-drop and 5–5–5 groups. In the 23-drop group, approximately 21 items were incorrectly recalled at first test; in the 5–5–5 group, approximately 10 items were incorrect at first test. Whereas the subset of 21 items still to be learned in the 23-drop group contained almost the full range of item difficulty, the 10 remaining items to be learned in the 5–5–5 group were likely the more difficult items. Thus, one would expect lower final test performance for a subset of relatively difficult items (as in the 5–5–5 group) than for a subset that also included some easier items (as in the 23-drop group).

Another interpretation of the results of the conditional analyses is that the advantage of the 23-drop schedule for initially incorrect items was due to the longer initial ISI (23 items) for this group in comparison with that of the 5–5–5 group (5 items). This longer ISI may have benefited final retention. Recent research suggests that longer ISIs lead to lower initial learning but greater final test performance than do shorter ISIs (see, e.g., Pashler et al., 2003).

Experiment 2 was designed to replicate the results of Experiment 1 and to address these alternative interpretations. First, given that this research represents the first investigation of the efficiency of variable schedules, replicating the advantage of the 23-drop schedule over a conventional schedule is important to establish that this outcome is robust. Accordingly, Experiment 2 included the 23-drop schedule and the 5–5–5 schedule from Experiment 1 in order to examine whether the 23-drop group again yielded similar performance in fewer trials and particularly benefited initially incorrect responses. Second, in order to address both of the alternative interpretations discussed above, we included a second conventional schedule (23–23–23) with an ISI of 23. Including this group in the design allowed comparison of the 23-drop schedule with a conventional schedule in which the initial ISI is equated. Additionally, the equivalent initial ISI should yield a similar number of incorrect responses on the initial test–retest trial, which would reduce concerns about differential item difficulty in the subset of initially incorrect items in each group.

Method

Participants and Design. Sixty-six undergraduate students enrolled in Introductory Psychology at Kent State University participated in return for course credit. Participants were randomly assigned to one of three groups: 5–5–5, 23–23–23, and 23-drop, with 21–23 participants in each group. Single or repeated test–retest trials (described below) was a within-subjects variable.

Materials and Procedure. Materials included the two lists of Swahili–English paired associates from Experiment 1. Twenty-four items were assigned to the repeated test–retest condition. In the 5–5–5 and 23–23–23 groups, each repeated item was given three test–retest trials. Initial study and each test–retest trial were separated by 5 or 23 intervening items, respectively. The schedule for the 23-drop group was the same as that in Experiment 1; 6 participants were affected by the stopping rule. Again, the lag in the dropout group varied after the first test–retest trial, depending on the number of items correctly recalled during a given trial. Across participants, mean lag between test–retest trials decreased from 23

between initial study and the first test–retest trial to 18, 12, 8, and 3 intervening items between subsequent test–retest trials.

In each group, the other 24 items were assigned to the single test–retest trial condition, with the assignment of list to condition counterbalanced across participants. These items were presented for an initial study trial followed by one test–retest trial (with the same ISI as that for the repeated items in each group). In comparison with the study-only condition that was used to provide filler items for the 1–5–9 schedule in Experiment 1, this control provides a better comparison condition to demonstrate the benefit of repeated test–retest trials. Appendix A contains more specific details about the item presentation schedule for each group. The procedure was otherwise identical to that in Experiment 1, including the final cued-recall test after a 40-min filled interval.

Results and Discussion

On the basis of the results of Experiment 1, we were primarily interested in whether the 23-drop schedule would yield similar performance levels in fewer trials than the conventional schedule groups. Mean number of test–retest trials used for the repeated items across participants in each group is reported in the top panel of Figure 2. Set values for the 5–5–5 and 23–23–23 groups are included in the figure for ease of comparison with the dropout group. Results from a one-sample *t* test indicated that the 23-drop group again used significantly fewer than 72 trials [$t(20) = 5.8, p < .001$].

Mean percentage of items correctly recalled at final test for each group is reported in the bottom panel of Figure 2. As in Experiment 1, performance for the repeated items is displayed in the bars of the figure; performance for the single test–retest control items is reported in parentheses within the bar for each group. A 2 (trials: repeated or single) \times 3 (schedule) mixed-factor ANOVA indicated a statistically significant main effect of trials [$F(1,63) = 96.2, p < .001$], indicating that repeated test–retest trials resulted in significantly higher final test performance than did a single test–retest trial. The main effect of schedule and the interaction were not significant [$F(2,63) = 2.69, p = .076$, and $F(2,63) = 2.06, p = .137$, respectively]. Although there was not a significant main effect of schedule, there was a numerical advantage for the repeated items in the 23-drop group as opposed to the two conventional groups. However, performance in the single test–retest condition was unexpectedly higher for the 23-drop group than it was for either the 5–5–5 or 23–23–23 groups. When performance in the single condition was used as a covariate, there was still no significant performance difference between groups in the repeated condition [$F(2,62) = 2.03, p = .14$]. Thus, taken together, the results indicate that all three groups attained similar levels of performance but that the dropout schedule did so in significantly fewer trials than the conventional schedules.

For interested readers, we again report the derived efficiency measure of recall per trial. Means (and standard errors) were .66 (.09), .72 (.08), and 1.30 (.15) for the 5–5–5, 23–23–23, and 23-drop groups, respectively. In contrast to Experiment 1, the recall per trial measure in Experiment 2 thus supports the same qualitative conclusion as the raw measures.

Final test performance as a function of the kind of response generated for an item on the first test–retest trial

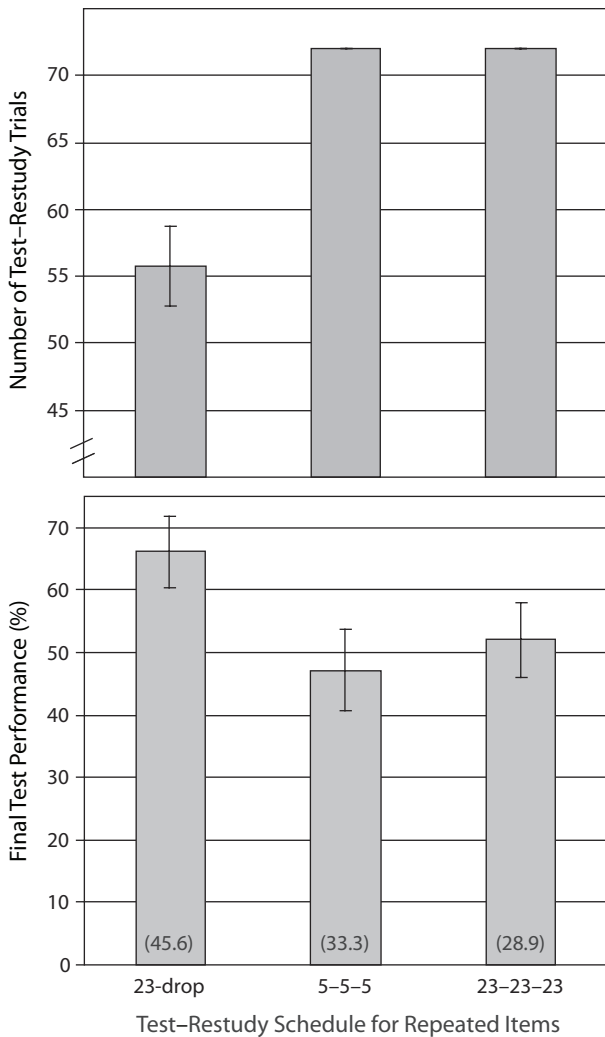


Figure 2. Top panel: Mean number of test–restudy trials used for repeated items across participants in each group in Experiment 2. Error bars represent standard errors. Number of trials was set at 72 in the 5–5–5 and 23–23–23 groups; these set values are presented for purposes of comparison with the 23-drop group. Bottom panel: Mean percentage of repeated items correctly recalled on final test across participants in each group in Experiment 2. Error bars represent standard errors. Values in parentheses represent percentage of control items correctly recalled on final test.

(either correct or incorrect) is reported in Table 2. Again, the comparison of the 23-drop schedule with each conventional schedule was of the greatest interest. First, consider final test performance for items that were correct on the initial test–restudy trial. Replicating the results of Experiment 1, final test performance for initially correct items was similar in the 23-drop and 5–5–5 groups, with a trend for an advantage in the 23-drop group [$t(41) = 1.83, p = .075$]. In contrast, final test performance for initially correct items was significantly lower in the 23-drop group than in the 23–23–23 group [$t(41) = 2.07, p = .044$]. Thus, when compared with a conventional schedule involving a longer ISI, the 23-drop group incurred some cost from discontinuing the practice of initially recalled items.

However, the cost of dropping recalled items from further practice was offset by gains from additional practice with initially incorrect items, which constituted the majority of initial responses. Initially incorrect items were significantly more likely to be correct at final test in the 23-drop group than in either the 5–5–5 group or the 23–23–23 group [$t(41) = 2.86, p = .007$ and $t(41) = 2.31, p = .026$, respectively]. Note that the 23-drop and 23–23–23 groups involved the same initial ISI and had a similar number of initially incorrect items. Thus, the advantage of the dropout schedule for initially incorrect items cannot be attributed to the length of the initial ISI and is unlikely to be due to differences in the difficulty of the incorrect items in the two groups. Rather, the most straightforward interpretation is that initially incorrect items benefited from the additional practice afforded by the dropout schedule. Consistent with this idea, on average, across participants in the 23-drop group, 4.7 of the initially incorrect items received four or more test–restudy trials (with some receiving up to six).

GENERAL DISCUSSION

Two experiments directly evaluated the efficiency of variable schedules of test–restudy practice—namely, dropout schedules—in comparison with conventional schedules. Results showed that a dropout schedule yielded levels of memory similar to those obtained with conventional schedules of learning and, importantly, did so using significantly fewer trials. Thus, dropout can be a more efficient schedule for studying to-be-learned materials.

Arguably as important, Experiment 1 demonstrated that not all dropout schedules will be more efficient than conventional schedules. Our results suggest that the overall effectiveness of dropout schedules may depend critically on the length of the interval between presentations of a given item. A dropout schedule with a longer initial ISI (the 23-drop group in Experiments 1–2) was more effective overall than conventional schedules, whereas a dropout schedule with a shorter initial ISI was not. Why was final test performance significantly lower in the 5-drop group (Experiment 1) than in the conventional schedule groups? One possibility is that individuals may have been covertly rehearsing items; thus, in the 5-drop group, some items may still have been readily available in working memory at the time of the initial test–restudy trial. If so, accessing the correct response from working memory would not necessarily be predictive of performance on a later test that required retrieval from long-term memory. Moreover, to the extent that retrieval from long-term memory leads to the strengthening of that information in long-term memory, items that were accessed from working memory at the time of the initial test–restudy trial rather than retrieved from long-term memory would not accrue further strength.

Another possible explanation for why final test performance was relatively low for initially correct items in the 5-drop group is that after an ISI of 5, these items had not proceeded far enough along the forgetting function to reach asymptote. If so, they would undergo further for-

Table 2
Final Test Performance As a Function of Kind of Response on First
Test–Restudy Trial for Repeated Items, Experiment 2

Group	Correct Recall on First Test–Restudy Trial				Incorrect Recall on First Test–Restudy Trial			
	No. Items Correctly Recalled		% Correct Recall on Final Test		No. Items Incorrectly Recalled		% Correct Recall on Final Test	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
5–5–5	17.3	0.89	51.2	6.7	6.7	0.89	37.8	7.8
23–23–23	5.6	0.94	84.7	6.1	18.4	0.94	46.2	6.1
23-drop	7.9	0.98	67.3	5.7	16.1	0.98	65.8	5.9

Note—Total number of items in each group was 24.

getting after the initial test–restudy trial. Consistent with this idea, Rubin, Hinton, and Wenzel (1999) showed that after only four intervening items, cued-recall performance for paired associates was still in the steeper part of the negatively accelerated short-term forgetting curve. In contrast, after 21 intervening items (similar to the ISI in the 23-drop group), performance was approaching asymptote (for similar results, see Pashler et al., 2003).

One might expect that in the 5-drop group, the test–restudy trial itself should have protected initially recalled items from further forgetting (referred to as *preventive maintenance*; Bahrick & Hall, 1991). Roediger and Karpicke (2006b) replotted data from Spitzer (1939) in order to illustrate the preventive maintenance function of testing. Results showed that administering a practice test significantly reduced forgetting across a delay in comparison with a condition in which no practice test was administered. Although the present research was not specifically designed to estimate preventive maintenance effects, the comparison of final test performance for initially correct items in the 5-drop group to overall final performance in the study-only control group in Experiment 1 (32% vs. 12%, respectively) is at least suggestive of the possibility that some preventive maintenance may have occurred. However, the extent of preventive maintenance was clearly less than that when the practice test occurred after a longer delay (67%–71%, as for the initially correct items in the 23-drop groups).

We note, however, that in Experiment 2, items that were correct on the first test–restudy trial were less likely to be correct at final test in the 23-drop group than they were in the 23–23–23 group. Thus, recallable items appeared to benefit from additional practice in the 23–23–23 group, presumably through further preventive maintenance that protected these items from forgetting over the retention interval. Interestingly, the benefit of additional practice for recallable items in a conventional schedule also appeared to depend critically on the length of the ISI—when additional practice trials for initially correct items were separated by a relatively short interval (5 items, as in the 5–5–5 group), those items were no more likely to be recalled on the final test than were items that were correctly recalled only once but after a longer interval (as in the 23-drop group). Perhaps most importantly, the small disadvantage for initially correct items in the 23-drop group

in comparison with the 23–23–23 group was completely offset by a sizeable gain in final test performance for items that were initially incorrect.

Although not the primary goal of the present research, our results are relevant to the debate about which of two conventional schedules (fixed interval or expanding interval, such as the 5–5–5 and 1–5–9 schedules used in Experiment 1) leads to higher levels of performance. Early research suggested that an expanding-interval schedule was the most effective conventional schedule (see, e.g., Landauer & Bjork, 1978), leading to higher levels of performance than did fixed-interval schedules. However, more recent studies have reported minimal differences between fixed-interval and expanding-interval schedules (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Carpenter & DeLosh, 2005; Cull, 2000). Our findings add to the growing body of literature suggesting that expanding-interval schedules are not more effective than fixed-interval schedules. These findings suggest that future research comparing conventional schedules with variable schedules of test–restudy need only include a fixed-interval schedule.

Overall, the present research provides initial evidence that a dropout schedule can be more efficient than the conventional schedules that have been widely studied in previous research. These initial findings are promising and provide a foundation for future research that can replicate and extend them in important directions. For example, we have begun comparing dropout and conventional schedules with more complex verbal materials (e.g., key ideas from expository text) than the paired associates used herein. Additionally, future research should investigate the effectiveness of dropout versus conventional schedules at longer retention intervals between practice and final test. Indeed, recent research using a free-recall paradigm suggests that examining the efficacy of dropout schedules for longer delays will be an important direction for further research (Karpicke & Roediger, 2007). Certainly, the efficiency of dropout schedules with more complex materials and longer test delays should be established before strong, general recommendations can be made for how students might best allocate their limited study time across various domains.

AUTHOR NOTE

The present research was supported by the Institute of Education Sciences, U.S. Department of Education Grant R305H050038, to Kent State

University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Thanks to John Dunlosky and the RADlab group for helpful discussion of this research and comments on earlier drafts of this article. Correspondence concerning this article should be sent to M. A. Pyc, Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001 (e-mail: mpyc@kent.edu).

REFERENCES

- BAHRICK, H. P. (1967). Relearning and the measurement of retention. *Journal of Verbal Learning & Verbal Behavior*, *6*, 89-94.
- BAHRICK, H. P., & HALL, L. K. (1991). Preventive and corrective maintenance of access to knowledge. *Applied Cognitive Psychology*, *5*, 1-18.
- 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 & Language*, *52*, 566-577.
- BALOTA, D. A., DUCHEK, J. M., SERGENT-MARSHALL, S. D., & ROEDIGER, H. L., III (2006). Does expanding retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology & Aging*, *21*, 19-31.
- CARPENTER, S. K., & DELOSH, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619-636.
- CARPENTER, S. K., PASHLER, H., & VUL, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826-830.
- CARRIER, M., & PASHLER, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633-642.
- 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.
- CULL, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215-235.
- CULL, W. L., SHAUGHNESSY, J. J., & ZECHMEISTER, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365-378.
- DRISKELL, J. E., WILLIS, R. P., & COPPER, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology*, *77*, 615-622.
- KARPICKE, J. D., & ROEDIGER, H. L., III (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory & Language*, *57*, 151-162.
- KRUEGER, W. F. C. (1929). The effect of overlearning on retention. *Journal of Experimental Psychology*, *12*, 71-78.
- 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: Academic Press.
- MORRIS, P. E., FRITZ, C. O., JACKSON, L., NICHOL, E., & ROBERTS, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology*, *19*, 779-798.
- NELSON, T. O., & DUNLOSKY, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*, 325-335.
- NELSON, T. O., LEONESIO, R. J., SHIMAMURA, A. P., LANDWEHR, R. F., & NARENS, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *8*, 279-288.
- PASHLER, H., CEPEDA, N. J., WIXTED, J. T., & ROHRER, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 3-8.
- PASHLER, H., ZAROW, G., & TRIPLETT, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning Memory & Cognition*, *29*, 1051-1057.
- ROEDIGER, H. L., III, & KARPICKE, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255.
- ROEDIGER, H. L., III, & KARPICKE, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181-210.
- ROHRER, D., TAYLOR, K., PASHLER, H., WIXTED, J., & CEPEDA, N. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, *19*, 361-374.
- RUBIN, D. C., HINTON, S., & WENZEL, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 1161-1176.
- SPITZER, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641-656.
- 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, & Cognition*, *25*, 1024-1037.
- WHEELER, M. A., EWERS, M., & BUONANNO, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571-580.

NOTES

1. The labels "restudy only" and "retrieval only" are used to refer to the explicit instructions in each condition. Of course, participants may have engaged in covert retrieval on some trials in the restudy-only condition; likewise, correct recall in the retrieval-only condition affords a restudy opportunity. Thus, the difference between these groups and the condition in which retrieval plus restudy was explicitly instructed on every trial may be a matter of degree.

2. There has been some debate in the literature about which conventional schedule (fixed interval or expanding interval) leads to higher performance levels (see General Discussion for elaboration on this topic). Given the mixed results in the literature (see, e.g., Cull, 2000; Landauer & Bjork), it was not clear which of these schedules should be used as the benchmark for comparison to dropout schedules. Thus, we included both a fixed and an expanding interval schedule in Experiment 1.

3. Unexpectedly, performance on the first test-restudy trial differed between the three groups with an initial ISI of 5, with more items initially correct in the 5-5-5 group than in either the 5-drop group or 5-drop-after-2 group. One explanation for performance differences between the groups at first test-restudy trial despite equivalent initial ISIs may be differential interference that was due to differences in the way items were blocked for test-restudy trials in each of the groups (see Appendix A). More specifically, the 5-5-5 group performed initial study and all test-restudy trials for a set of items before moving onto the next set, whereas both dropout groups only performed initial study and one test-restudy trial with a given set of items before being introduced to the next set of items. In the 5-5-5 group, finishing all practice with a set of items at one time may have resulted in less proactive interference with subsequent items. If so, the interference explanation would predict similar levels of performance between groups for the first set of items (items 1-6 in Appendix A), with differences emerging only for later sets of items. Results were consistent with this account: Mean performance for the first set of items for 5-5-5, 5-drop, and 5-drop-after-2 were 4.2, 3.9, and 3.7, respectively, with no significant differences between groups (all $t_s < 0.81$). In contrast, performance was significantly greater for the 5-5-5 group than for the 5-drop and 5-drop-after-2 groups for the second set of items (5.1, 4.1, and 4.1, respectively, $t_s > 2.16$) and for the third set of items (5.3, 4.0, and 4.5, respectively, $t_s > 1.79$). Thus, it appears that the overall differences at first practice test for the three groups with an initial ISI of 5 were at least partially due to differential interference for intermediate sets of items.

APPENDIX A
Overview of Presentation Schedule for Control and
Experimental Items in Each Group for Experiments 1 and 2

Experiment 1

5-5-5

- Initial study for control items 1-12
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 1-6
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 7-12
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 13-18
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 19-24
- Initial study for control items 13-24

1-5-9

- Study of control items 1-24 interspersed throughout initial study and test-restudy trials for repeated items 1-24

5-drop

- Initial study for control items 1-12
- Initial study, test-restudy for repeated items 1-6
- Initial study, test-restudy for repeated items 7-12
- Initial study, test-restudy for repeated items 13-18
- Initial study, test-restudy for repeated items 19-24
- Continue test-restudy trials for any remaining repeated items
- Initial study for control items 13-24

5-drop-after-2

- Initial study for control items 1-12
- Initial study, test-restudy for repeated items 1-6
- Initial study, test-restudy for repeated 7-12
- Initial study, test-restudy for repeated 13-18
- Initial study, test-restudy for repeated items 19-24
- Test-restudy repeated items 1-24
- Test-restudy remaining repeated items
- Study control items 13-24

23-drop

- Initial study for control items 1-12
- Initial study, test-restudy for repeated items 19-24
- Continue test-restudy trials for any remaining repeated items
- Initial study for control items 13-24

Experiment 2

5-5-5

- Initial study, test-restudy for control items 1-6 here or at end*
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 1-6
- Initial study, test-restudy for control items 7-12
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 7-12
- Initial study, test-restudy for control items 13-18
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 13-18
- Initial study, test-restudy for control items 19-24
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 19-24

23-23-23

- Initial study, test-restudy for control items 1-24 here or at end*
- Initial study, test-restudy, test-restudy, test-restudy for repeated items 1-24

23-drop

- Initial study, test-restudy for control items 1-24 here or at end*
- Initial study, test-restudy for repeated items 1-24
- Continue test-restudy trials for any remaining repeated items

* Presentation of these items at the beginning or end of the practice list was counterbalanced across participants.

APPENDIX B
Number of Items Tested on Each Test–Restudy Trial for Repeated Items During Practice,
the Percentage of Those Items Tested on a Given Test–Restudy Trial That Were Correctly
Recalled on That Trial, and the Number of Repeated Items That Were Not Correctly Recalled
on Any Test–Restudy Trial During Practice

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5		No. Items Incorrectly Recalled at Least Once in Practice	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
	Experiment 1											
5-drop	46.0	3.5	20.9	3.4	39.3	5.8	41.0	5.7	53.4	9.2	3.7	1.0
Number of items tested	24.0	0	13.0	0.8	10.9	0.9	7.2	0.8	2.4	0.4		
23-drop	14.3	2.5	24.6	3.3	37.9	3.5	55.7	5.3	63.9	11.0	7.0	1.4
Number of items tested	24.0	0	20.6	0.6	16.0	1.0	5.1	0.5	0.8	0.2		
5-drop-after-2	45.2	3.4	33.6	3.2	29.8	3.8	45.0	5.0	87.5	12.5	6.2	1.0
Number of items tested	24.0	0	24.0	3.0	17.6	0.7	5.2	0.5	0.2	0.1		
5–5–5	59.0	3.6	81.1	3.6	89.5	3.0					1.7	.7
Number of items tested	24.0	0	24.0	0	24.0	0						
1–5–9	68.6	3.5	68.0	4.3	70.8	3.9					0	0
Number of items tested	24.0	0	24.0	0	24.0	0						
Experiment 2												
23-drop	32.7	4.1	41.7	5.1	46.2	6.3	61.7	7.1	83.3	11.8	3.1	1.2
Number of items tested	24.0	0	16.1	1.0	10.2	1.2	4.7	0.9	0.7	0.2		
23–23–23	23.2	3.9	42.4	4.9	56.3	6.1					9.0	1.5
Number of items tested	24.0	0	24.0	0	24.0	0						
5–5–5	72.2	3.7	88.1	1.7	95.1	1.6					0.2	0.1
Number of items tested	24.0	0	24.0	0	24.0	0						

(Manuscript received November 9, 2006;
revision accepted for publication April 2, 2007.)