Applied Cognitive Psychology, Appl. Cognit. Psychol. **25**: 87–95 (2011)
Published online 23 November 2009 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.1646

Costs and Benefits of Dropout Schedules of Test–Restudy Practice: Implications for Student Learning

MARY A. PYC* and KATHERINE A. RAWSON

Kent State University, USA

Summary: Almost all previous studies examining the benefits of testing for promoting student learning have used fixed schedules of practice. However, students more often report utilizing a dropout schedule of practice, in which items are dropped from practice once they are known. Two experiments investigated the costs and benefits of utilizing a dropout schedule of test–restudy practice. Participants learned Swahili–English paired associates using a dropout schedule or a fixed schedule. In the dropout schedule, items received test–restudy practice until each item was correctly recalled once. In the fixed schedule, all items received three tests–restudy practice trials regardless of whether they were correctly recalled, as in previous research. Experiment 2 also included a second learning session. In both experiments, a final cued recall test was administered several days later. Results indicated that the benefits of the dropout schedule (fewer practice trials used overall and all items correctly recalled once during practice) need to be considered in light of the costs (lower levels of final test performance). Copyright © 2009 John Wiley & Sons, Ltd.

Previous research has shown that multiple distributed test–restudy trials are effective for promoting memory (e.g. Carpenter & DeLosh, 2005; Cull, 2000; Cull, Shaughnessy, & Zechmeister, 1996; see also Roediger & Karpicke, 2006 for a recent review). However, most previous research on schedules of test–restudy practice has used fixed schedules of practice in which all items are practiced a predetermined number of times (usually three test–restudy trials for each item), regardless of whether they are correctly or incorrectly recalled during practice. In contrast, students more often report using *dropout* schedules of practice, in which items are dropped from practice after a variable number of practice trials at different times throughout practice (e.g. Kornell & Bjork, 2007, 2008). In a survey study by Kornell and Bjork (2008), 56% of students reported using flashcards to study (akin to the laboratory method in which students are prompted with a cue and attempt to retrieve a target). Of those students, approximately 75% reported using a dropout method, in which they placed flashcards to the side once an item was correctly recalled during practice. Furthermore, nearly all previous research on testing effects has involved only one learning session (but see Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), and no previous study evaluating the efficacy of dropout versus fixed schedules has involved more than one learning session. However, students presumably study during more than one learning session, particularly when using flashcards.

How might the efficacy of dropout schedules of test–restudy practice compare to the fixed schedules examined in previous research, particularly with more than one learning session? The main goal of the present research was twofold. The first goal was to understand the costs and benefits of the dropout method of test–restudy practice that is more commonly used by students than the fixed practice schedules that have been examined in previous research. The second goal was to understand the costs and benefits of dropout versus fixed schedules of practice when learning occurs during more than one session.

Concerning the potential benefits of dropout schedules of practice, consider the time demands on college students studying for final exams. Not only does each exam require memory for a large amount of material, but students often have multiple exams to study for at the same time. Thus, students have a large amount of information they need to know, but a limited amount of time in which to learn it. Therefore, effective students must study course material enough to be able to remember it later, but they must also be economical (or efficient) in the amount of time they continue studying material they know to allow sufficient time to study material they have not yet learned well. To this end, dropout schedules of test–restudy practice may be more efficient for students than fixed schedules of practice if the same level of performance can be achieved in fewer practice trials. Another potential benefit of dropout schedules compared to fixed schedules of practice concerns the likelihood that all items will be learned during practice. Specifically, students typically practice items until they have been correctly recalled at least once before dropping them from practice. In contrast, fixed schedules involve a set number of practice trials for each item, and thus practice may be terminated in a fixed schedule before all items have been learned.

Regarding the potential benefits of a dropout schedule of practice, Pyc and Rawson (2007) compared dropout and fixed schedules of test–restudy practice using Swahili–English paired associates. In the *dropout* schedule, each item received test–restudy practice until it was correctly recalled once during practice. In the *fixed* schedule, each item received three practice test–restudy trials regardless of whether or not it was correctly recalled during practice. Dropout and fixed schedules led to similar levels of performance on a final cued recall test administered after a 40-minute delay. However, the dropout schedule was more efficient, in that fewer practice trials (approximately 2.3 test–restudy trials per item) produced the same level of

*Correspondence to: Mary A. Pyc, Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA.
E-mail: mpyc@kent.edu

performance as the fixed schedule in which the number of practice trials was set at three trials per item (as in much of the prior research on repeated test–restudy practice).

As in much of the previous research on retrieval practice, Pyc and Rawson (2007) examined performance after a relatively short retention interval. However, students are generally expected to retain information across longer delays. What are the costs and benefits of the dropout method of test–restudy practice when assessed using longer retention intervals, which arguably are of greater relevance for education?

Karpicke and Roediger (2007a, Experiment 2) recently reported evidence that the benefits of dropout schedules must be weighed against their significant costs at longer retention intervals. Of interest here, they presented participants with a list of unrelated words for test–restudy practice, either using a dropout schedule in which test and restudy were dropped for an item after it was correctly recalled once during practice or a fixed schedule in which each item received eight test–restudy trials. On a final free recall test one week later, recall was significantly lower in the dropout condition than in the fixed condition.

Similarly, Karpicke and Roediger (2008) compared dropout and fixed schedules of practice using lists of Swahili–English paired associates. Of interest here, significantly fewer trials were used during practice in the dropout schedule compared to the fixed schedule. Additionally, dropout and fixed schedules of learning showed similar learning rates and absolute learning levels during practice. However, on the final cued recall test one week later, performance was significantly lower in the dropout condition than in the fixed condition.

Finally, Kornell and Bjork (2008, Experiment 1) also compared dropout and fixed schedules using a one week retention interval. In their dropout condition, participants had up to 10 minutes to practice items and could drop items from test–restudy practice at their discretion; in the fixed condition, items continued to be presented for test–restudy practice until 10 minutes had elapsed. On the final test 1 week later, performance was significantly lower in the dropout condition than in the fixed condition.

In sum, relatively few studies have examined the costs and benefits of dropout versus fixed schedules of test–restudy practice. Although initial work suggested that the benefits of the dropout schedule are greater than the costs (Pyc & Rawson, 2007), recent work with longer retention intervals appears to point to the opposite conclusion. However, methodological differences exist between these studies that may have influenced the relative advantage of dropout versus fixed schedules (see Appendix for a detailed outline of each of the experiments discussed above). First, items in the fixed schedules used by Karpicke and Roediger (2007a) received eight test–restudy trials (versus the three test–restudy trials more typical of research using fixed schedules), drastically exaggerating the difference in amount of practice in the two groups. Additionally, practice tests involved free recall as opposed to cued recall, as in most other studies. Second, Kornell and Bjork (2008) allowed participants to decide when to drop items from practice, and thus some items may have been dropped from practice before they were correctly recalled at least once, particularly because students were only given 10 minutes to practice items. Third, items in Karpicke and Roediger (2008) were studied in a blocked fashion (all items tested and then all items restudied), which is dissimilar from how students would test and study items when using flashcards. Fourth and most important, no previous study has included a second learning session (to foreshadow, in Experiment 2 we consider theoretical reasons for why a second learning session may influence the pattern of performance for dropout vs. fixed schedules).

Accordingly, we conducted two experiments to further investigate the costs and benefits of a dropout schedule compared to a fixed schedule with a longer retention interval. In Experiment 1, we used the same basic method as Pyc and Rawson (2007) but with a one week retention interval. In Experiment 2, we extended beyond all previous research exploring the costs and benefits of dropout schedules by implementing a second learning session two days after the initial learning session. Students who use flashcards to study presumably use them to self-test during more than one learning session. Thus, in keeping with our highest-level goal of investigating the schedules of self-testing that are more typical of those students use, Experiment 2 examined the costs and benefits of a dropout schedule that involved a second learning session.

EXPERIMENT 1

Method

Participants and design

Twenty-two undergraduate students enrolled in Introductory Psychology at Kent State University participated in return for course credit.[1] Schedule of practice (dropout or fixed) was a within-participant variable.

Materials and procedure

Materials included 48 Swahili–English translation word pairs, divided into two lists with an equivalent range of item difficulty on each word list (based on norms reported by Nelson & Dunlosky, 1994). Assignment of list to practice schedule condition was counterbalanced across participants. All task instructions and items were presented via computer. Half of the participants completed the dropout schedule first followed by the fixed schedule, and the other half completed the two schedules in the opposite order.

In the *fixed* practice schedule, each of the 24 items received an initial study trial in which the Swahili word was presented on the left side of the screen and the target English translation appeared on the right side of the screen for 10 seconds. After all items were presented for initial study,

[1] Five additional participants completed the study but learned fewer than half of their dropout items to criterion (one correct recall during practice) and thus were dropped from analyses (because these participants were likely not adhering to instructions). Note that dropping these individuals from the analyses does not contaminate comparisons of the efficacy of dropout versus fixed schedules of practice because schedule of practice was a within-participant manipulation. On average for the remaining participants, 0.8 items (SE = 0.4) did not reach criterion during practice in the dropout schedule.

each item then received three test–restudy practice trials. For each test–restudy practice trial, the Swahili word was presented alone and participants had Eight seconds to enter the English translation in a text box. After eight seconds, the response box was removed from the screen, and the Swahili and English words were presented together for four seconds of restudy. We used an interstimulus interval (ISI) of 23 items between initial study and each subsequent test–restudy practice trial for each item.

In the *dropout* condition, all items first received an initial study trial, as in the fixed condition. All items then received one test–restudy practice trial, with 23 items intervening between initial study and the first test–restudy practice trial for each item. If the correct English translation was recalled, the item was dropped from further practice (*cf.* a student setting aside a flashcard for an item that was correctly recalled). If the translation was not correctly recalled, the item was placed at the end of the list of to-be-learned items for another test–restudy practice trial (*cf.* a student placing a flashcard for an incorrectly recalled item at the back of the stack of cards). This process continued until either all items were correctly recalled once or a participant reached a 72-trial maximum allowance (not including the initial study trials). The maximum allowance was set at 72 as in Pyc and Rawson (2007) so that the amount of practice in the dropout schedule did not exceed the 72 test–restudy practice trials allotted in the fixed schedule. On average, 0.8 items (SE = 0.4) did not reach criterion during practice in the dropout schedule.

Upon completion of the learning session, participants were dismissed and reminded to return for the second session one week later. The final test was a participant-paced cued-recall test, with one item presented at a time via computer. Note that this procedure is an exact replication of Pyc and Rawson (2007) with the exception that the retention interval was one week here versus 40 minutes in the earlier study.

Results and discussion

We conducted a preliminary analysis of final test performance as a function of order of schedule of practice (dropout or fixed schedule completed first). Results indicated no significant difference in final test performance, so all further analyses collapse across this variable.

Regarding the benefits of using a dropout schedule of practice, we computed the total number of practice trials used in the dropout schedule. On average, participants used 54.9 test–restudy trials (SE = 3.0), compared to the set 72 trials used in the fixed schedule of practice. A one-sample *t*-test revealed that the dropout condition used significantly fewer than 72 trials, $t(21) = 5.71, p < .001$, Cohen's $d = 1.72$. Thus, the dropout schedule was more economical in the number of practice trials used during practice compared to the fixed schedule.

Figure 1 reports the cumulative proportion of items correctly recalled for each trial during practice. On one hand, the learning rates over the first three trials were similar for fixed and dropout schedules of practice. On the other hand, the absolute learning level achieved at the end of the learning session for these two schedules was clearly different. By the end of the session, 97% of the items in the dropout schedule had been correctly recalled, whereas only 77% of the items in the fixed schedule had been correctly recalled at least once during practice. Thus, one benefit of dropout schedules of practice is that a higher percentage of items are correctly recalled by the end of practice in dropout compared to fixed schedules of practice.

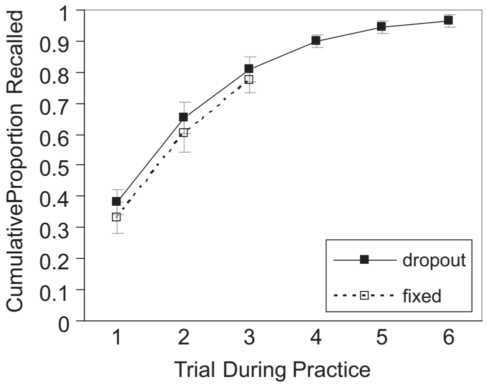

Figure 1. Cumulative proportion of words that were correctly recalled on or before each practice trial during the learning phase, for each practice schedule condition in Experiment 1

However, on the final test, the mean percentage of items correctly recalled was significantly lower in the dropout versus fixed condition 16.0 (SE = 2.9) and 34.3 (SE = 5.0), respectively, $t(21) = 4.30, p < .001, d = .95$. Thus, although the dropout schedule used fewer trials overall than the fixed schedule, the economy of using fewer trials during practice must be weighed against the lower level of final test performance in the dropout condition.

To further examine the trade-off between the number of practice trials used during practice and final test performance, we computed a derived measure of efficiency. For each participant in each condition, we divided final test performance by the number of trials used during practice to yield the gain in final test performance (in percentage points) from each practice trial (hereafter referred to as *gain per trial*). Across participants, gain per trial was lower in the dropout condition (.37, SE = .08) than in the fixed condition (.48, SE = .07), $t(21) = 1.61, p = .122, d = .30$. Although not significant, this trend suggests that the costs of the dropout schedule are greater than the benefits, at least as compared to a fixed schedule of practice.

Why was final test performance lower in the dropout condition? One possibility is that it was due to a limit in the number of times an item could be correctly recalled during practice. To explore this possibility, we conducted conditional analyses. For each individual, we examined final test performance as a function of the first trial in which an item was correctly recalled during practice (Tulving, 1964). Means across individuals for each condition are reported in Figure 2 (C = correct, N = not correct; values in boxes represent the mean number of items contributing to each condition). We made three comparisons between dropout and fixed conditions based on the first trial in which an item was correctly recalled during practice. Comparing the two practice conditions for subsets of items based on the first trial in which an item was correctly recalled allows

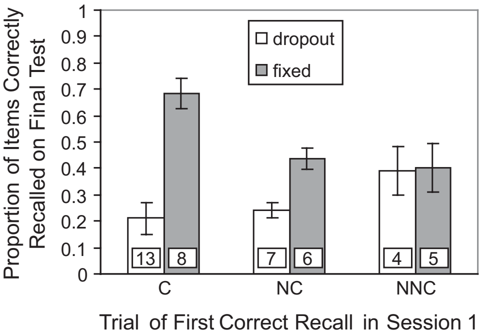

Figure 2. Proportion of items recalled at final test as a function of the first correct recall for individual items during the learning phase in Session 1. Values in boxes represent the number of items contributing to means for each comparison

examination of the influence of additional practice after one correct recall during learning while providing some control for item difficulty (e.g. in both practice conditions, items correctly recalled on the first trial were presumably the easiest items).

First, we compared final test performance for items that were initially correctly recalled on the first practice trial in dropout and fixed conditions (C items in Figure 2). Necessarily, items in the dropout condition were dropped from practice after this first correct recall, whereas the subset of items in the fixed condition could be correctly recalled up to two more times during practice (on average, in the fixed condition, C items were correctly recalled 2.7 times during practice).[2] Final test performance for C items was significantly lower in the dropout condition than in the fixed condition, $t(10) = 5.58$, $p < .001$, $d = 1.94$. A similar pattern was observed for items that were initially correctly recalled on the second trial (NC items in Figure 2) during practice (on average in the fixed condition, NC items were correctly recalled 1.9 times during practice). Final test performance for NC items was significantly lower in the dropout condition than in the fixed condition, $t(9) = 3.45$, $p = .007$, $d = 1.45$. Finally, we compared performance for items that were initially correctly recalled on the third practice trial (NNC items in Figure 2). Note that these items were only ever correctly recalled one time during practice in both the fixed and dropout schedules of practice. Results showed no significant difference in final test performance for the NNC items in the dropout and fixed conditions, $t(3) = .30$, $p = .785$, $d = .23$.[3]

Taken together, these results suggest that overall performance was lower in the dropout versus fixed condition due to a limit on the number of times items could be correctly recalled. Another non-exclusive explanation concerns the functional lag between trials in the two conditions. In the fixed condition, the lag between practice trials for a given item was constant throughout practice (because no items dropped out). In the dropout condition, as easier items were correctly recalled and dropped from practice, fewer items remained to maintain the lag between trials for the harder items. However, the NNC item results suggest that a contracting lag contributed minimally to the lower level of final test performance in the dropout condition. That is, NNC items in the dropout schedule were learned with a contracting lag, whereas NNC items in the fixed schedule were not, but final test performance was nonetheless similar for these items.

To summarize the overall pattern of results, the dropout schedule was beneficial to the extent that fewer trials were used during practice and almost every item was correctly recalled once during practice. However, the dropout schedule yielded lower final test performance overall, replicating results from previous studies using long retention intervals but different methods. Results from the conditional analyses suggest that the cost to final test performance may be due to a limit on the number of times items are correctly recalled during practice. These conditional analyses are new to studies examining the costs and benefits of dropout schedules of practice and further clarify why the dropout schedule of practice is at a disadvantage in terms of final test performance compared to the fixed schedule of practice.

EXPERIMENT 2

Results of Experiment 1 indicate that discontinuing practice after items have been correctly recalled only once leads to lower final test performance, compared to a fixed schedule in which items could be correctly recalled up to three times during practice. Thus, students may experience limited gains from self-testing with flashcards to the extent that they drop items after only one correct recall during a learning session (the most common study decision reported by students in the survey by Kornell & Bjork, 2008). However, students who study with flashcards presumably use them to self-test during more than one learning session. Thus, it is important to evaluate the efficacy of schedules of practice with more than one learning session, which has not yet been investigated in relation to the efficacy of dropout versus fixed schedules of practice.

How might dropout schedules compare to fixed schedules with a second learning session, which is presumably more similar to how students spontaneously schedule their flashcard practice? Here, we briefly consider two theoretical accounts that support different predictions concerning dropout versus fixed schedules of learning after a second learning session.

First, the strategy shift hypothesis (Bahrick & Hall, 2005) predicts that a second learning session may be more beneficial for a dropout schedule compared to a fixed schedule of practice. Broadly speaking, the strategy shift hypothesis states that retrieval failures encourage learners to

[2] The number of times items were correctly recalled for C items in the fixed condition was less than three because the C subset includes CCC, CCN, CNC and CNN items. 97% of C items were CCC items, and results are qualitatively similar when analyses only include those items. Similarly, in analyses of NC items to follow, the number of times items were correctly recalled for NC items in the fixed condition was less than two because the NC subset includes NCC and NCN items. 95% of NC items were NCC items, and results are similar with just these items.

[3] We realize that the degrees of freedom are quite low in some cases because not all participants contributed values to each cell. Although some caution is warranted in interpreting these results, they are nonetheless suggestive. Additionally, to foreshadow, similar patterns were observed in Experiment 2 with a somewhat larger number of participants contributing to each comparison.

shift to using a more effective encoding strategy during subsequent learning. More effective strategies will be better for later retrieval, such as during a final retention test (e.g. Richardson, 1998; Dunlosky & Hertzog, 2000). Although the experimental method used here constrains the global-level strategy (test-restudy practice), the specific encoding strategy used (particularly during restudy) is up to participants. For example, participants may shift from using less effective to more effective encoding strategies (e.g. from rote repetition to interactive imagery or the keyword method; Bahrick & Hall, 2005). Additionally, participants may adapt the kind of mediator used within a given strategy (e.g., non-interactive vs. interactive imagery). Results of Experiment 1 suggest that on the first test trial of a second learning session, the dropout condition will likely recall fewer items compared to the fixed condition. According to the strategy shift hypothesis, more retrieval failures may actually be beneficial if participants shift to using effective strategies for items that were incorrectly recalled. Thus, the dropout schedule may perform similarly to a fixed schedule at final test because better strategies have been developed for items that were initially incorrectly recalled during the second learning session.

Alternatively, the desirable difficulty framework (Bjork, 1994) predicts that even after a second learning session, the disadvantage in final test performance for the dropout schedule versus the fixed schedule will persist (and if anything will increase). Broadly speaking, the desirable difficulty framework states that difficult but successful processing is most beneficial for memory (as compared to easy successful processing or difficult unsuccessful processing). Previous research has shown that the difficulty of a correct retrieval increases as lag between trials increases (Karpicke & Roediger, 2007b; Pyc & Rawson, 2009). As related to the current experiment, it follows that correctly recalling an item on the first trial of the second learning session (after two days have elapsed) will be more difficult than correct recalls that come later during that session (e.g., 1–2 minutes after a restudy opportunity). Based on results from Experiment 1, more items will be correctly recalled on the first test trial during a second learning session in a fixed schedule compared to a dropout schedule of practice. Therefore, the desirable difficulty framework predicts that the fixed schedule will continue to outperform the dropout schedule at final test because a greater proportion of items in the fixed schedule are correctly recalled with more difficulty on the first test trial during the second learning session compared to the dropout schedule.

Note that the primary motivation for inclusion of the second learning session in Experiment 2 was not to competitively evaluate these two theoretical positions, although the outcomes may indirectly bear on these accounts. Rather, we discuss these theoretical frameworks here to illustrate the plausibility of alternative empirical outcomes.

Experiment 2 examined the costs and benefits of a dropout schedule that involved a second learning session. The method in Experiment 2 was the same as Experiment 1, except that participants again practiced items during a second learning session two days after Session 1.

Method

Participants and design

Twenty-eight undergraduate students enrolled in Introductory Psychology at Kent State University participated in return for course credit.[4] Schedule of practice (dropout or fixed) was a within-participant variable.

Materials and procedure

The lists of Swahili–English paired associates, assignment of list to practice schedule condition, and Session 1 procedure were the same as in Experiment 1. Session 2 was administered two days later. The only procedural difference between Sessions 1 and 2 was that the initial study trial was eliminated from Session 2; thus, the first practice test trial of Session 2 could also serve as a criterion test. For the fixed schedule, all items received a restudy trial after the initial test trial and then two additional test–restudy practice trials. For the dropout schedule, all items received a restudy trial after the initial test trial, regardless of whether they were correctly or incorrectly recalled, but items correctly recalled on the initial trial were then dropped from further test–restudy practice. Items incorrectly recalled on the initial test trial continued to be tested and restudied until they were correctly recalled once during practice, as in Session 1, until all items were correctly recalled once or the 72-trial maximum allowance was reached. On average, 1.0 ($SE = 0.5$) and 0.1 ($SE = 0.1$) items did not reach criterion during practice in the dropout schedule during Session 1 and Session 2, respectively. Upon completion of Session 2, participants were dismissed and reminded to return for the third session one week later. The final test during Session 3 was a participant paced cued-recall test, as in Experiment 1.

Results and discussion

We conducted a preliminary analysis of final test performance as a function of order of schedule of practice (dropout or fixed schedule completed first). Results indicated no significant difference in final test performance, so all further analyses collapse across this variable.

Regarding the benefits of a dropout schedule of practice, participants in the dropout condition used 54.5 trials ($SE = 2.2$) during Session 1 and 46.9 trials ($SE = 1.9$) during Session 2. One-sample t-tests revealed that the dropout condition used significantly fewer than 72 trials in both sessions, $t(27) = 8.00$, $p < .001$, $d = 2.1$ and $t(27) = 13.16$, $p < .001$, $d = 3.5$. Thus, the dropout schedule was again more economical in the number of practice trials used during practice compared to the fixed schedule.

Figure 3 reports the cumulative proportion of items correctly recalled for each trial in Session 1 (Panel A) and Session 2 (Panel B). As in Experiment 1, the learning rate for the two practice schedules was similar but the absolute learning level achieved by the end of the learning session was not. By the end of Session 1, 95% of the items in the dropout

[4] Nine additional participants completed the study. However these participants learned fewer than half of their dropout items to criterion in Session 1 and thus were dropped from analyses.

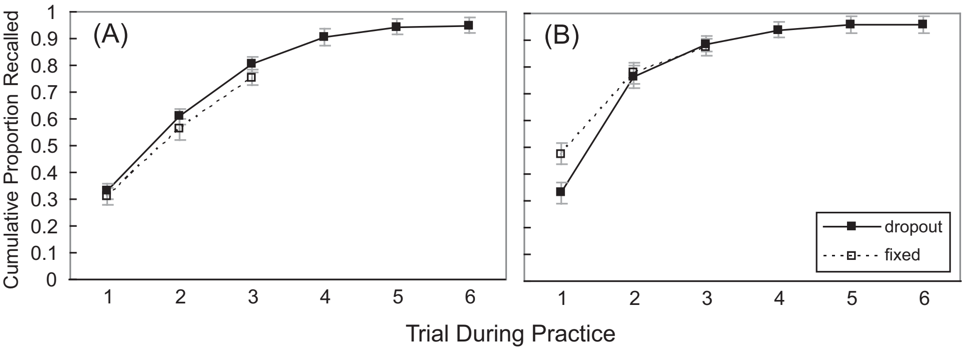

Figure 3. Cumulative proportions of words that were correctly recalled on or before each practice trial during the learning phase in Session 1 (Panel A) and Session 2 (Panel B), for each practice schedule condition in Experiment 2

schedule had been correctly recalled, whereas only 75% of the items in the fixed schedule had been correctly recalled at least once. By the end of Session 2, 96% of the dropout items in the dropout schedule had been correctly recalled, whereas 88% of the items in the fixed schedule had been correctly recalled at least once. Thus, again we found that one benefit of dropout schedules is that more items are correctly recalled by the end of a given learning session than with a fixed schedule.

Regarding the costs of using a dropout schedule of practice, we examined test performance at two time points, the initial practice test trial in Session 2 (which can be treated as a criterion test, given that no restudy was provided prior to that test trial in Session 2) and the final test in Session 3. The mean percentage of items correctly recalled on the first practice test trial in Session 2 in the dropout and fixed conditions is reported in Panel A of Figure 4. Performance was significantly lower in the dropout condition than in the fixed condition, $t(27) = 3.51$, $p = .002$, $d = .60$, which mirrors the results from Experiment 1.

Of greater interest, were the costs associated with the dropout schedule maintained even after a second learning session? The mean percentage of items correctly recalled at final test in each condition are reported in Panel B of Figure 4. Once again, final test performance was lower for a dropout schedule compared to a fixed schedule, $t(27) = 3.42$, $p = .002$, $d = .59$, even when learning involved a second session, with no hint of the dropout schedule catching up (interaction term from a 2 x 2 repeated measures ANOVA, $F < 1$). Experiment 2 provides further evidence that the economy of using fewer trials in a dropout schedule must be weighed against the decrease in final test performance relative to a fixed schedule.

To further examine the trade-off between the number of practice trials used during learning and final test performance, we calculated the derived efficiency measure of gain per trial (as in Experiment 1) for each learning session. In Session 1, mean gain per trial was .67 (SE = .09) in the dropout condition versus .65 (SE = .06) in the fixed condition, $t(27) = .30$, $p = .766$, $d = .07$. In Session 2, mean gain per trial was 1.31 (SE = .14) in the dropout condition versus .95 (SE = .05) in the fixed condition, $t(27) = 3.14$, $p = .004$, $d = .63$. In contrast to the results of Experiment 1, gain per trial was similar or better in the dropout versus fixed condition in Experiment 2. However, the advantage in derived efficiency for the dropout condition must be considered in light of the overall lower level of performance in this condition, which tempers any prescriptions concerning the efficacy of the dropout schedule.

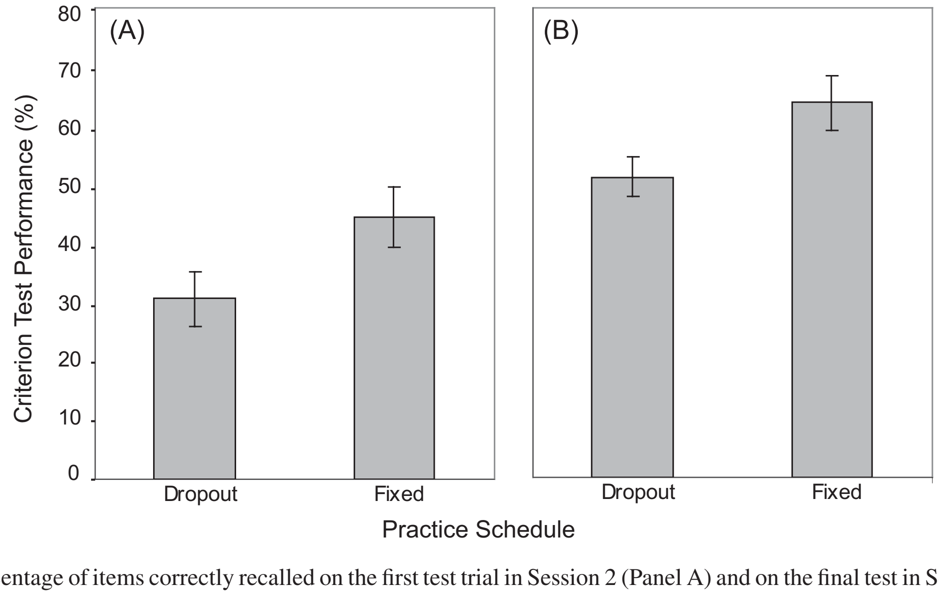

Figure 4. Mean percentage of items correctly recalled on the first test trial in Session 2 (Panel A) and on the final test in Session 3 (Panel B), for each practice schedule condition in Experiment 2. Error bars represent standard errors

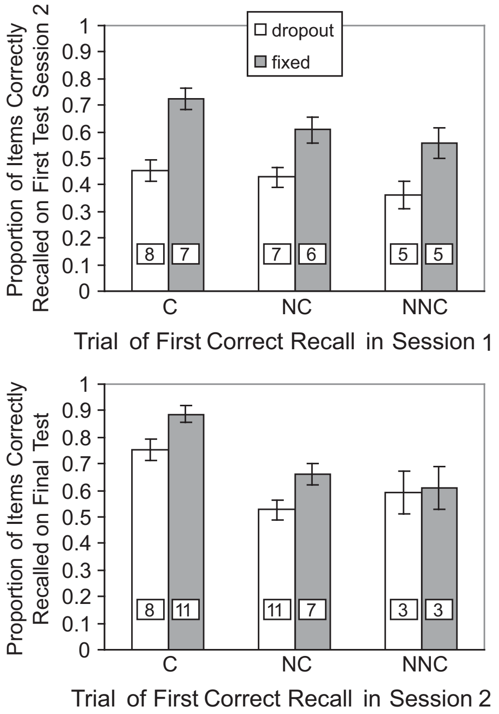

Figure 5. Proportion of items recalled at first test in Session 2 (top) and final test (bottom) as a function of the first correct recall for individual items during the learning phase in Session 1 (top) and Session 2 (bottom). Values in boxes represent the number of items contributing to means for each comparison

To examine the extent to which differences in final test performance for the dropout schedule versus the fixed schedule were due to differences in the number of times items were correctly recalled during practice, we conducted conditional analyses similar to those in Experiment 1. First, we examined performance on the first test trial in Session 2 as a function of the first trial in which an item was correctly recalled in Session 1 (top panel, Figure 5). As we would expect based on results from Experiment 1, performance was lower for items correctly recalled only once during Session 1 in the dropout schedule versus items that were correctly recalled multiple times in the fixed schedule [C items, $t(25) = 5.35$, $p < .001$, $d = 1.2$ and NC items, $t(21) = 3.68$, $p = .001$, $d = .93$]; on average in the fixed condition, C items were correctly recalled 2.9 times (SE = .04) and NC items were correctly recalled 1.9 (SE = .02) times during the learning session. Finally, performance on the first test trial in Session 2 for NNC items was lower for the dropout versus fixed condition, $t(13) = 2.59$, $p = .022$, $d = .73$, which suggests that contracting lag may also have contributed to some extent to the overall performance disadvantage for the dropout condition.

Second, we examined final test performance as a function of the first trial in which an item was correctly recalled in Session 2 (bottom panel, Figure 5). The pattern of results mirrors those from Experiment 1. Final test performance was lower for items correctly recalled only once in Session 2 in the dropout schedule versus items that were correctly recalled multiple times in the fixed schedule [C items, $t(26) = 2.66$, $p = .013$, $d = .73$ and NC items, $t(25) = 2.25$, $p = .033$, $d = .59$]; on average in the fixed condition, C items were correctly recalled 2.8 times (SE = .06) and NC items were correctly recalled 1.9 (SE = .04) times during Session 2.[5] Additionally, final test performance was similar for NNC items in the dropout and fixed conditions, $t(8) = 1.06$, $p = .318$, $d = .24$. Overall, the pattern of results converges with those from Experiment 1, providing further evidence that students may be limiting the efficacy of self-testing by discontinuing practice with items after only one correct recall during a given learning session. Furthermore, this disadvantage persists even when items in the dropout condition are relearned to criterion during a second learning session.

GENERAL DISCUSSION

The main goal of the present experiments was to evaluate the costs and benefits of dropout methods of test–restudy practice that are commonly used by students, when final performance is assessed after a longer retention interval and particularly when more than one learning session is involved. On one hand, the dropout schedule was beneficial in that it consistently used significantly fewer practice trials than the fixed schedule used in much of the previous research on repeated retrieval practice, an important issue when considering the time constraints students are faced with when studying for multiple exams. On the other hand, the absolute savings in time was minimal (3–5 minutes, given 12 seconds per trial). Furthermore, the costs of using a dropout schedule for subsequent memory are significant. The dropout schedule consistently produced lower levels of final test performance compared to the fixed schedule of practice. Thus, the benefits accrued from utilizing fewer test–restudy trials must be weighed against the significant disadvantage in final test performance.

On one hand, our results and conclusions converge with those from recent research examining the efficacy of dropout versus fixed schedules of practice with a one week retention interval (Karpicke & Roediger, 2007a; Karpicke & Roediger, 2008; Kornell & Bjork, 2008). On the other hand, the present work provides important extensions beyond these earlier studies. Most important, Experiment 2 provides a unique contribution to the literature evaluating the efficacy of dropout versus fixed schedules of practice by implementing a second learning session, which is presumably more typical of the schedules used by students who self-test with flashcards. Additionally, the current experiments extend beyond earlier studies by evaluating final test performance as a function of the first time an item was correctly recalled during a given learning session to explore the relation between the number of correct recalls during practice and later test performance, particularly when learning involves more than one session.

The current findings diverge from those reported by Pyc and Rawson (2007) who used a much shorter retention interval. The different pattern of results in these two studies is consistent with previous testing effect research showing

[5] In Session 1, 86% of C items were CCC items and 93% of NC items were NCC items; in Session 2, 95% of C items were CCC items and 94% of NC items were NCC items. Results are qualitatively similar when analyses only include those items.

that schedules of learning that are relatively effective at one retention interval may be less effective at other retention intervals (e.g., lag effects at short versus long retention intervals, Pyc & Rawson, 2009). This pattern also mirrors results found in related literatures. In the spacing effect literature, massed study often leads to similar or higher levels of final test performance than distributed study with short retention intervals (Bloom & Shuell, 1981; Cull, 2000; Toppino & Gracen, 1985). In contrast, distributed study typically leads to higher levels of final test performance than massed study with longer retention intervals (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Rawson & Kintsch, 2005). This reversal in the efficacy of distributed versus massed schedules of practice as a function of retention interval is sometimes referred to as Peterson's paradox, after results reported by Peterson, Wampler, Kirkpatrick, and Saltzman (1963; see also Balota, Duchek, & Paullin, 1989 for a more recent replication of these results).

Taken together, these findings have important implications for student learning. Kornell and Bjork (2008) reported that 75% of the students who used flashcards to study reported using a dropout method, and students most frequently discontinued practice with items after only one correct recall. Thus, students may be aware of the potential benefits for utilizing a dropout schedule (economy) but may not be aware of the costs of such a schedule (lower test performance due to discontinuing practice too early). The present results indicate that these costs are likely to persist even if students self-test to a criterion of one correct recall in more than one learning session. Regarding general prescriptive conclusions from the present results, however, we would advise against throwing the baby out with the bathwater. Although dropping items after one correct recall appears to be sub-optimal for memory after longer retention intervals, dropout schedules may be more beneficial if items are not dropped after the first correct recall. Additionally, dropout schedules may be more beneficial than fixed schedules for review sessions that occur shortly before an exam.

In addition to applied implications, the current results also have some theoretical implications. The results from Experiment 2 were in line with predictions from the desirable difficulty framework (Bjork, 1994), clearly showing that the fixed schedule of practice continued to outperform the dropout schedule even when a second learning session occurred. The prediction from the desirable difficulty framework is based on the claim that greater difficulty correctly recalling items will be better for memory. Because more items were correctly recalled on the first trial in Session 2 (for the fixed schedule compared to the dropout schedule of practice), more items were correctly recalled with greater difficulty, which presumably in turn yielded a persistent advantage for the fixed schedule at final test. Thus, these results add to growing support for the desirable difficulty framework.

In sum, the present study bridges the gap between the fixed schedules of test–restudy practice most commonly studied in prior research and the understudied dropout schedules of practice that students more commonly report using. These results can guide both future research and recommendations to students about how to more effectively regulate their study.

ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants #R305H050038 and #R305A080316 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 Jill Peterson for assistance with data collection. Special thanks also to John Dunlosky and RADlab for helpful comments and feedback on this project.

REFERENCES

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., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*, 3–9.

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.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*, 245–248.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636.

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.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

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.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging*, *15*, 462–474.

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

Karpicke, J. D., & Roediger, H. L. III. (2007b). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719.

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

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

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

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

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology*, *66*, 206–209.

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

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.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology*, *97*, 70–80.

Richardson, J. T. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review*, *5*, 597–614.

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.

Toppino, T. C., & Gracen, T. F. (1985). The lag effect and differential organization theory: Nine failures to replicate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 185–191.

Tulving, E. (1964). Intratrial and intertribal retention: Notes towards a theory of free recall verbal learning. *Psychological Review*, *71*, 219–237.