MEMORY, 2008, 16 (2), 125–136

Psychology Press
Taylor & Francis Group

Optimising self-regulated study: The benefits—and costs—of dropping flashcards

Nate Kornell and Robert A. Bjork
University of California, Los Angeles, CA, USA

Self-regulation of study activities is a constant in the lives of students—who must decide what to study, when to study, how long to study, and by what method to study. We investigated self-regulation in the context of a common study method: flashcards. In four experiments we examined the basis and effectiveness of a metacognitive strategy adopted almost universally by students: setting aside (dropping) items they think they know. Dropping has a compelling logic—it creates additional opportunities to study *un*dropped items—but it rests on two shaky foundations: students' metacognitive monitoring and the value they assign to further study. In fact, being allowed to drop flashcards had small but consistently negative effects on learning. The results suggest that the effectiveness of self-regulated study depends on both the accuracy of metacognitive monitoring and the learner's understanding, or lack thereof, of how people learn.

Self-regulated learning involves any number of decisions, such as whether one has memorised a word pair, or mastered Rachmaninov's notoriously difficult piano concerto number 3, and whether to test oneself or not. Research on self-regulated learning has mainly focused on two variables: the amount of time people spend on a given item, and the likelihood that they will choose to study an item at all (Metcalfe & Kornell, 2005). The current experiments represent an attempt to investigate another common study decision—whether it is time to *stop* studying.

Perhaps no memorisation technique is more widely used than flashcards, especially during homework. When people study with flashcards, they often "drop"—that is, put aside and stop studying—items they think they know. Dropping items that seem well learned has a compelling logic: It creates more opportunities for the remaining items to be studied and, in fact, the best-selling flashcards available on the internet (a set of GRE flashcards) is specially designed and marketed to encourage dropping.

How effective, though, is the dropping strategy? One potential problem is that dropping items relies on metacognitive monitoring, which can be flawed, as well as one's understanding, or lack thereof, of the value of future study opportunities. Another is that dropping items changes the subsequent sequencing of events, including the spacing of repetitions of items that are not dropped. The goal of the flashcard-inspired experiments we report is to clarify the memory and metamemory processes and consequences that characterise self-regulated study.

METAMEMORY CONSIDERATIONS IN SELF-REGULATED STUDY

Self-regulated study relies on two basic aspects of metacognition: making judgements about one's

Address correspondence to: Nate Kornell, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095-1563, USA. E-mail: nkornell@psych.ucla.edu

Grant 29192G from the McDonnell Foundation supported this research. We thank Bridgid Finn, Matt Hays, Jason Finley, and Dan Fink for their help in running the experiments, and Mark A. McDaniel and Robert M. Nosofsky for their comments on an earlier draft of this article.

http://www.psypress.com/memory
DOI:10.1080/09658210701763899

learning and memory (monitoring) and using those judgements to guide study behaviour (control) (Nelson & Narens, 1994). Errors in either aspect can lead to ineffective study decisions.

Metacognitive monitoring

Metacognitive judgements are made based on a variety of cues (e.g., Koriat, 1997), such as retrieval fluency (e.g., Benjamin, Bjork & Schwartz, 1998; Kelley & Lindsay, 1993), cue familiarity (e.g., Metcalfe, Schwartz & Joaquim, 1993; Reder & Ritter, 1992), and the success (or lack thereof) of previous retrieval attempts (e.g., Dunlosky & Nelson, 1992; Spellman & Bjork, 1992).

The accuracy of metacognitive monitoring is measured in two ways—by *resolution,* which is high if better-learned items are given relatively high ratings, and *calibration*, which is high to the degree that people's predicted recall levels match their actual recall levels. Both types of monitoring accuracy can affect study choices: Poor resolution can lead people to prioritise the wrong items; poor calibration, particularly overconfidence, can lead to too much dropping and too little studying.

The type of monitoring required in the current experiments—judgements of learning (JOLs)—can be unreliable, but when participants are allowed (or required) to test themselves, JOLs have been shown to be quite accurate—in terms of both resolution (e.g., Dunlosky & Nelson, 1992) and calibration (e.g., Koriat, Ma'ayan, Sheffer, & Bjork, 2006). With respect to dropping an item, therefore, a virtue of flashcards is that self-testing is intrinsic to the test/study nature of flashcard practice.

Metacognitive control

Metacognitive monitoring is useful only if coupled with an effective control strategy. That is, learners must decide, given their monitoring, which items will profit most from additional study. The Region of Proximal Learning (RPL) model of study-time allocation, for example, says that learners should give priority to items that are close to being learned, not those already learned, or those too difficult to learn (e.g., Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2003). In the context of flashcards, people adopting the RPL strategy should drop cards they think they already know, as well as cards they believe they cannot learn.

Whether or not to drop an item is a deceptively complex decision. With a fixed amount of time to study, dropping an item leaves more time for the remaining items to be studied, but any particular item can always be dropped the next time around. Participants must therefore decide which has more value: studying the current item one additional time (at least), or dropping it in favour of preserving one (or more) additional opportunity to study some other item before the end of the allotted time. According to the RPL idea, the value of studying is highest for items that are closest to being learned, making it critical to guard against dropping items too quickly.

A student who focuses too much on learning the most difficult flashcards is in danger of dropping easier items too soon. Study decisions depend on a student's goals (Dunlosky & Theide, 1998). Unlike students who set easily achievable goals—who tend to choose relatively easy materials to study (Dunlosky & Thiede, 2004; Thiede & Dunlosky, 1999)—students who believe they can master all of the too-be-learned materials typically focus on the most difficult materials (see Son & Metcalfe, 2000, for a review). A strong focus on difficult flashcards translates to a strong desire to drop easy flashcards—even if doing so means jeopardising one's ability to recall the easy ones later.

MEMORY CONSIDERATIONS

In addition to the perils of selecting an effective dropping strategy, there are a number of drawbacks to dropping items from study that students may not be aware of. One is that spacing, as opposed to massing, study opportunities on a given item has been shown many times to have large benefits for memory (e.g., Cepeda, Pashler, Vul, Wixted & Rohrer, 2006). Dropping items has the possible drawback that it decreases the spacing of the repetitions of the remaining items. Participants are unlikely to appreciate this subtlety, given that they sometimes rate spaced practice as less effective than massed practice (e.g., Baddeley & Longman, 1978; Simon & Bjork, 2001).

A second consideration is that dropping also undermines the positive effects of overlearning—that is, continuing to study an item one already knows (Christina & Bjork, 1991; but see also

Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005, for evidence that the effects of overlearning diminish with time). Karpicke and Roediger (2007) have shown that there are tremendous benefits to restudying something one already knows, if the restudy takes the form of a test. In their experiment, after participants answered an item correctly once it was dropped, either from future presentations or from future tests. Dropping correct items from future presentations had negligible effects, but dropping them from future tests had dramatic and negative effects on long-term retention. Because flashcards involve testing, the implication of such findings is that dropping flashcards after only one successful recall attempt might be an extraordinarily bad strategy.

In Karpicke and Roediger's (2007) research, it was virtually impossible for dropping items to have positive effects, because an item being dropped simply resulted in less study time overall. In research by Pyc and Rawson (in press), on the other hand, dropping one item allowed participants to spend more time on other items (as in the experiments reported here). In that situation, equivalent learning was achieved in the drop and no-drop conditions, but the drop condition required less study time, implying that dropping has value. However, in Pyc and Rawson's (in press) study, dropping was controlled by a computer not the participants. Before concluding that students should drop flashcards when they study, it is important to examine the effect of self-regulated dropping.

In summary, then, there are good reasons to expect the intuitive promise of dropping flashcards to be coupled with some actual benefits (e.g., Pyc & Rawson, in press), especially given that allowing people to decide how they study has had positive results in previous experiments (e.g., Kornell & Metcalfe, 2006; Nelson, Dunlosky, Graf, & Narens, 1994). However, there are also reasons why dropping items might be perilous and problematic. The experiments we report were designed to clarify the memory and metamemory processes that are intrinsic to self-regulated study and the consequences of those processes for learning.

EXPERIMENT 1

Participants studied two lists of English–Swahili translations for 10 minutes each. The procedure was similar to studying flashcards: Participants cycled through the same cards repeatedly, and on each trial the front of the card was shown first, allowing the participant to test himself or herself, before the card appeared to flip and the back was shown. Participants were allowed to drop items while studying one of the two lists (Drop condition), but not the other (No-drop condition). A cued-recall test on all of the words was administered either immediately or after a week's delay.

Method

Participants. A total of 60 Columbia University students participated during one of four lab sessions to fulfil a class requirement. There were 31 and 29 participants in the immediate and delayed conditions, respectively.

Materials. The materials were 40 English–Swahili translations, 20 per list, selected from a set published by Nelson and Dunlosky (1994). Each list contained a mixture of easy (e.g., cloud-wingu), medium (e.g., lung-pafu), and difficult (e.g., forgery-ubini) pairs.

Design. The experiment was a 2 (Study-control: Drop vs No-drop) × 2 (Delay: Immediate vs Delayed) mixed design, with Study-control and Delay manipulated within and between participants, respectively. The order of the Drop and No-drop lists was counterbalanced across participants.

Procedure. The instructions described the experiment as similar to studying with flashcards, and explained the procedure in detail. Each of the two lists was then presented for 10 minutes, and participants were allowed to study as many items as they could in that time. A clock at the top right corner of the screen counted down the time remaining for study.

The translations were presented one word at a time. First the "front" of the card (the English cue) was shown for 1.5 s; then the card appeared to flip, and the "back" of the card (the Swahili target) appeared for 3 s. After the target disappeared, participants in the Drop condition were asked to choose—by selecting either a "Study again later" button or a "Remove from stack" button—whether to keep the item in the stack during subsequent cycles through the list (i.e., put it at the back of the "stack"), or drop it. In the No-drop condition, only the "Study again later" button was presented. Participants were prompted to hurry if they took longer than 4 seconds to make their choice.

If a participant dropped all of the word pairs, the screen remained blank until 10 minutes was up. This aspect of the procedure, which was explained in the instructions, was necessary to equate the time spent on the Drop and No-drop lists, and also discouraged participants from trying to hasten the end of the experiment by dropping all of their cards.

During the final-test phase, the words from the two lists were mixed and tested in random order. The English cue was shown and participants were asked to type in the Swahili target. Participants were prompted to hurry if they took more than 12 seconds to respond.

Results and discussion

During the study phase, for participants in both the Immediate and Delayed conditions, the average number of times an item was presented was higher in the No-drop condition than in the Drop condition, (5.29 vs 4.60 and 5.29 vs 4.92, respectively; *SD*s = .37, 1.06, .26, .62, respectively). (The distributions in each of the four conditions were negatively skewed, because many participants reached close to the maximum possible number of study trials.) The effect of Study-control was significant, $F(1, 58) = 21.41$, $p < .0001$, $MSE = .39$, $\eta_p^2 = .27$. The effect of Delay condition on number of study trials was not significant, $F(1, 58) = 1.53$, $p = .22$, $MSE = .47$, $\eta_p^2 = .026$, nor was the interaction, $F(1,58) = 1.96$, $p = .17$, $MSE = .39$, $\eta_p^2 = .033$. An average of 14.52 ($SD = 6.54$) and 13.79 ($SD = 7.60$) items were dropped from study in the Immediate and Delayed conditions, respectively, a difference that was not significant, $t(58) = .40$, $p = .69$. (In both conditions, the distribution was characterised by a large number of participants who dropped the maximum possible number of items).

Participants did not benefit from being allowed to control their study. On the contrary, as Figure 1 shows, test accuracy was significantly worse in the Drop condition than the No-drop condition, $F(1, 58) = 9.97$, $p < .01$, $MSE = .020$, $\eta_p^2 = .15$. Not surprisingly, performance was better on the Immediate test than on the Delayed test, $F(1, 58) = 56.88$, $p < .0001$, $MSE = .089$, $\eta_p^2 = .50$, but delay did not interact with the study-control manipulation, $F(1, 58) = .58$, $p = .45$, $MSE = .020$, $\eta_p^2 = .010$. In a separate analysis we excluded participants who, in their drop-condition list, dropped all of the pairs before 10 minutes had elapsed (12 and 14 participants were excluded in the Immediate and Delayed conditions, respectively, leaving 19 and 15 participants in those conditions). Final test accuracy remained better in the No-drop condition ($M = .36$, $SD = .30$) than the Drop condition ($M = .33$, $SD = .31$), but the effect was no longer significant, $F(1, 32) = .89$, $p = .35$, $MSE = .018$, $\eta_p^2 = .026$. We return to this point in the General Discussion. There was also a significant effect of delay, $F(1, 32) = 40.42$, $p < .0001$, $MSE = .077$, $\eta_p^2 = .56$, but no significant interaction, $F(1, 32) = .41$, $p = .52$, $MSE = .018$, $\eta_p^2 = .013$.

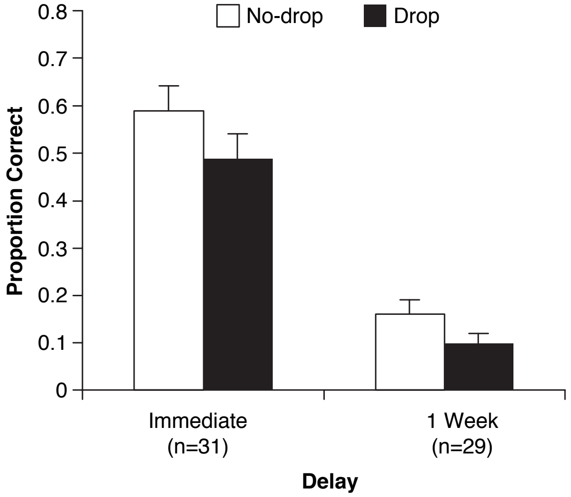

Figure 1. Proportion correct on the final test in Experiment 1 as a function of Study-control condition and test delay. Error bars represent standard errors.

EXPERIMENT 2

Experiment 2 was designed to explore the relative contributions of poor metacognitive monitoring and bad study strategies to the negative effect of self-regulation obtained in Experiment 1. With respect to monitoring, it seemed possible that overconfidence led participants in Experiment 1 to drop items sooner than they should have. To explore this possibility, half of the participants in Experiment 2 were asked to make a judgement of learning (JOL) whenever they dropped an item from study. We also explored whether participants had ill-conceived study strategies via a questionnaire administered at the end of the experiment.

Method

Participants and materials. The participants were 112 UCLA undergraduates who participated for course credit. There were 54 and 58 participants in the JOL and No-JOL conditions, respectively. The materials were the same as in Experiment 1.

Design. The experiment was a 2 (Study-control: Drop vs No-drop) × 2 (JOL condition: JOL vs No-JOL) mixed design. Participants in the JOL group were asked to make a JOL immediately after choosing to drop an item in the Drop condition; participants in the No-JOL group were not asked to make JOLs in the Drop condition.

Procedure. The procedure was similar to Experiment 1, but with four changes. First, participants in the JOL group were asked to make a JOL each time they dropped an item in the Drop condition. They did so—when prompted by "Chance you'll remember that one on the test"—by selecting one of six buttons, which were labelled 0%, 20%, 40%, 60%, 80%, and 100%. Second, the Swahili word became the cue and the English word the target, making the task easier. Third, to ensure that participants had time to test themselves during study, the cue was shown for 3 s (instead of 1.5 s). Finally, on the final test, all of the pairs from the first list were tested in random order, followed by all of the pairs from the second list.

Between the study and test phases, there was a 5-minute distractor task, during which participants were asked to identify famous people based on photographs presented upside-down. A post-experimental questionnaire asked participants a series of questions about their experience in the experiment and their study habits outside the laboratory.

Results and discussion

Study phase. During the study phase, in both the JOL and No-JOL conditions, the average number of times an item was presented was higher in the No-Drop condition than in the Drop condition (4.14 vs 3.80 and 4.15 vs 3.80, respectively; *SD*s = .23, .69, .21, .75, respectively). (The distributions were negatively skewed, as in Experiment 1.) This difference, collapsed over JOL condition, was significant, $F(1, 110) = 27.40$, $p < .0001$, $MSE = .25$, $\eta_p^2 = .20$. As the numbers make clear, the difference did not interact with group, $F(1, 110) = 0$, $p = .99$, $MSE = .25$, $\eta_p^2 = 0$, nor was the effect of making JOLs significant in the Drop condition, $F(1, 110) = .001$, $p = .97$, $MSE = .52$, $\eta_p^2 = 0$. An average of 14.01 ($SD = 6.88$) and 12.13 ($SD = 6.77$) items were dropped from study by the No-JOL and JOL groups, respectively, a difference that was not significant, $t(110) = -1.52$, $p = .13$. (In both conditions, a relatively large number of participants dropped all 20 items, as in Experiment 1.)

Final recall. Figure 2 shows the proportion of items correctly recalled by the No-JOL group (left panel) and JOL group (right panel). Consistent with the results of Experiment 1, the trend was towards impaired learning when participants were allowed to control their study (combined over the JOL and No-JOL groups), although the effect size was small and the effect was only marginally significant, $F(1, 110) = 3.01$, $p = .086$, $MSE = .027$, $\eta_p^2 = .026$. There was no overall main effect of JOL condition, $F(1, 110) = .88$, $p = .35$, $MSE = .13$, $\eta_p^2 = .008$, nor did JOL group interact with Study-control, $F(1, 110) = .092$, $p = .76$, $MSE = .027$, $\eta_p^2 = 0$.

Of the 112 participants, 37 (14 of 54 in the JOL group and 23 of 58 in the No-JOL group) dropped all of their pairs before 10 minutes had elapsed in the Drop condition. When the data were reanalysed including only participants who did not drop all of their pairs, final test performance

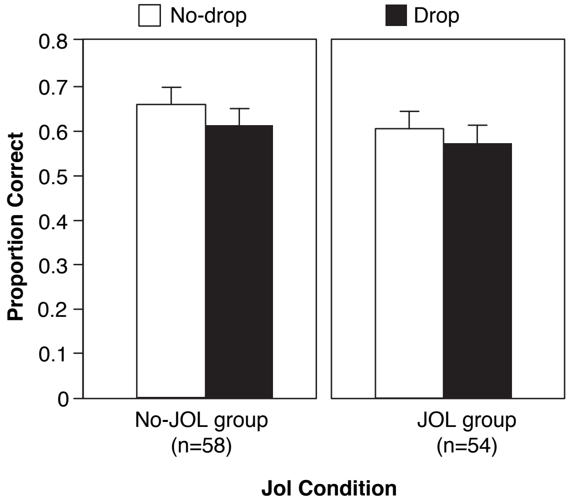

Figure 2. Proportion correct on the final test in Experiment 2 as a function of Study-control condition and JOL group. JOLs were only made in one condition, the JOL/Drop condition, represented by the rightmost bar.

remained better in the No-drop condition ($M = .61$, $SD = .27$) than the Drop condition ($M = .59$, $SD = .26$), but the effect was no longer significant, $F(1, 73) = .26$, $p = .61$, $MSE = .028$, $\eta_p^2 = .003$. The effect of JOL condition was not significant, $F(1, 73) = .35$, $p = .55$, $MSE = .11$, $\eta_p^2 = .005$, nor was the interaction, $F(1, 73) = .002$, $p = .97$, $MSE = .028$, $\eta_p^2 = 0$. Whether participants who dropped all items should be included or excluded from the analysis is discussed in the General Discussion.

Judgements of learning. Participants in the JOL group were quite accurate in predicting the likelihood that they would be able to recall the items they decided to drop. JOLs averaged 51% ($SD = 25$) and recall of items on which JOLs were made averaged 56% ($SD = 39$), a small under-confidence effect that was not significant, $t(51) = 1.14$, $p = .26$. Participants' JOLs were also accurate in terms of resolution: The average Gamma correlation between JOLs and test accuracy was significantly greater than zero, $M = .59$ ($SD = .51$), $t(26) = 6.03$, $p < .0001$ (given how Gamma is calculated, only the 27 participants with at least one correct and one incorrect response, at a minimum of two JOL levels, could be analysed). Overall, then, the fact that participants did not learn more when they were allowed to regulate their study (in the Drop condition) than when they were not (in the No-drop condition) appears attributable to factors other than poor metacognitive monitoring, as measured by calibration or resolution.

A possible contributor to the JOL participants' high levels of metacognitive accuracy is that they reported having tested themselves while studying, which increases both resolution (e.g., Dunlosky & Nelson, 1992) and calibration (e.g., Koriat et al., 2006). A total of 80% of participants said "yes" in response to the question "While you were studying, did you try to retrieve the English word on the 'back' of the card while you were looking at the Swahili word (on the 'front' of the card)?"

Surprisingly, the distribution of participants' JOLs for dropped items was roughly normal, as shown in Figure 3. This distribution is strikingly at odds not only with our prior expectations, but also with participants' self-reported study strategies. In response to the question "What made you decide to drop a word from your stack (instead of keeping it)?", 79% of participants reported dropping items that were easy or items that they felt they had learned. The remaining participants reported dropping the hardest items (17%) or a mixture of easy and hard items (4%). Given that pattern, one might have expected the most frequent responses to be JOLs of 100 (corresponding to items that participants perceived as already learned), followed by JOLs of 0 (corresponding to items perceived as too hard)—but 0 and 100 were the least frequent responses.

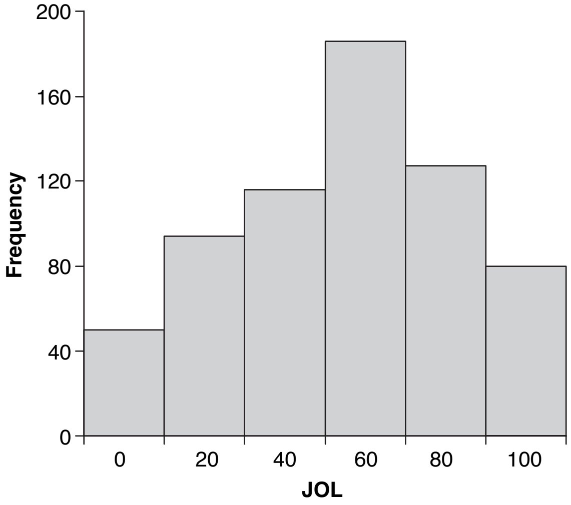

Figure 3. Frequency of responses at each JOL level in Experiment 2. JOLs were made immediately, and only, after a given item was dropped.

A possible interpretation of the distribution of JOLs, given participants' self reports, is that they adopted the strategy of studying a given item until they knew it *now*—that is, on the tests embedded in each cycle through the flashcards—even if they thought they might not remember it on the final test. Thus, as in other studies, the participants apparently under-weighted the positive consequences of additional study (see Koriat, Sheffer, & Ma'ayan, 2002; Kornell & Bjork, 2006; Rohrer et al., 2005) and self-tests (which appear to be especially important for items that can already be recalled; see Karpicke & Roediger, 2007).

Final recall revisited. If participants' metacognitive monitoring was accurate, then perhaps their failure to benefit from dropping items was caused by ineffective study strategies. We analysed each of the two general categories of self-reported study strategies separately. When the 79% of participants who reported dropping easy/known items were analysed, average accuracy was identical in the Drop and No-drop conditions ($M = .62$, $SD = .30$ in both cases). Thus, while not effective, these participants' study strategies were not harmful. The 21% of participants who reported dropping either the hard items or a mixture of hard and learned items, by contrast, contributed heavily to the negative effect of dropping items. Their test performance in the Drop condition ($M = .52$, $SD = .17$) was significantly lower than in the No-drop condition ($M = .70$, $SD = .22$), $F(1, 19) = 9.49$, $p < .01$, $MSE = .029$, $\eta_p^2 = .33$.[1] There was no main effect of JOL condition, $F(1, 19) = .69$, $p = .42$, $MSE = .053$, $\eta_p^2 = .035$, nor was there an interaction, $F(1, 19) = .003$, $p = .96$, $MSE = .029$, $\eta_p^2 = 0$. For these participants in particular, the hypothesis that poor study strategies contributed to the negative effect of dropping was supported. We discuss the RPL model in light of this finding in the General Discussion.

Post-experimental questionnaire. On the post-experimental questionnaire participants were asked "Do you study with flashcards in real life? If so, do you remove cards from your stack as you go?" In response, 56% said they study with flashcards and, of those, 75% said they dropped items as they studied.

EXPERIMENT 3

The results of Experiment 2 suggested that participants failed to profit from being able to drop items because they failed to appreciate the benefits of continuing to study an item after they could recall the target correctly. Experiment 3 was designed to explicitly examine the number of times participants recalled a target correctly before deciding to drop a given pair.

Method

Participants and materials. The participants were 25 UCLA undergraduates who participated for course credit. The materials were the same as in Experiments 1 and 2.

Procedure. The procedure was similar to Experiment 2 with two exceptions: both lists were assigned to the Drop condition and after the Swahili cue word was presented for 3 s, participants were asked to type in the English response word. A 3-s presentation of the correct English word followed, after which the participant chose to drop the pair or keep it in the list for further study. The first time through each list participants were not asked to type in responses, because they had yet to be exposed to the correct answers.

Results and discussion

In order to analyse the data conservatively with respect to the hypothesis that participants drop items too quickly, we treated misspelled answers as correct (because participants may have believed their answers were correct when they decided to drop), and we included only the 22 participants who reported using a strategy of dropping easy or known items (the other three participants frequently dropped items that they had never answered correctly).

Figure 4 shows the percentage of items that were dropped after zero, one, two, three, and four correct responses, pooled across lists and participants (no participant answered correctly five times or more). The data in Figure 4 represent all trials on which an item was dropped in either list, pooled across participants. The majority of items were dropped after one correct response. Surprisingly, 13% of the items were dropped after no correct responses, despite the fact that all of the participants included in the analysis reported dropping easy or known items. A hindsight bias

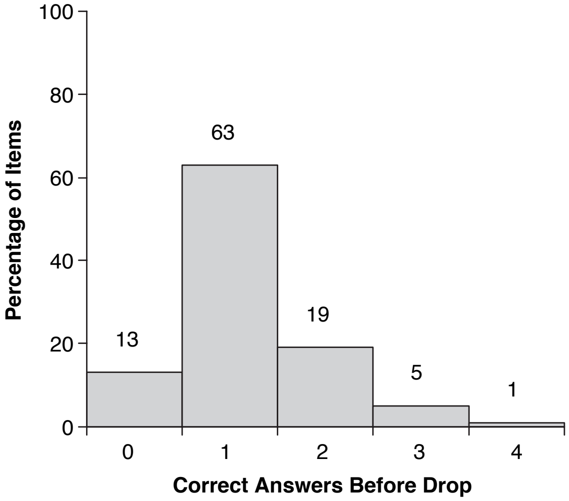

Figure 4. Frequency of dropping after zero, one, two, three, and four correct responses in Experiment 3.

[1] Participants who did not drop any items and participants who did not give an interpretable answer to the strategy question on the questionnaire were excluded from this analysis. When all 32 participants who did not report dropping easy/known items are included, the effect remains significant, $F(1, 30) = 10.47, p < .01, \eta_p^2 = .26$. The effect also remains significant when the two participants from this group who dropped all of the items in less than 10 minutes are excluded, $F(1, 28) = 7.44, p < .05, \eta_p^2 = .21$.

may have caused participants to believe that they had actually known the answer after it was shown, even though they could not recall it when tested. In total, 75% of the items that were dropped were dropped after less than two successful recall attempts.[2]

The finding that people dropped items very quickly, usually after a single correct recall attempt, may help to explain why participants in Experiments 1 and 2 did not benefit from being allowed to control their study. The data support that conclusion; final test accuracy for items dropped after zero, one, or two correct responses was .19, .62, and .88 ($SD = .15$, .24, .17, respectively).[3] Waiting to recall an item twice before dropping it increased final test performance by 26 percentage points compared to recalling it once, and 69 percentage points compared to not recalling it at all. Had participants waited to drop an item until they had recalled it more times, it appears as though they might have benefited from dropping. The drawback of waiting to drop an item, however, is that other items, which have not been dropped, receive less extra attention and may be learned less well as a result. In Experiment 4 we examined how different dropping strategies affect all items, by controlling the number of times an item was recalled before it was dropped.

There is a compelling, if counterproductive, logic to terminating study after one successful recall attempt. After a first successful recall, future recall success is almost guaranteed in the short term (e.g., Landauer & Bjork, 1978). Given that people seem to think of tests primarily as diagnoses of memory, not as learning events (Kornell & Bjork, 2007a; Kornell & Son, 2006), they may, paradoxically, think that there is little point in studying, or testing oneself on, an item that has been successfully recalled. That is, participants may reason as follows: There is no point in returning to an item that I will surely get correct next time anyway, and if I got it this time, I will surely get it next time, so why not drop the item? The flaw in this logic, of course, is that restudying an item that one can already retrieve correctly can have enormous memory benefits (e.g., Karpicke & Roediger, 2007; Landauer & Bjork, 1978).

EXPERIMENT 4

The results of Experiment 3 suggested that, in the first two experiments, dropping flashcards was ineffective because participants did so too eagerly, usually after a single correct recall. Experiment 4 tested that hypothesis. Each participant completed a Drop list and a No-drop list, but in three between-participant conditions, items in the drop list were dropped either (a) automatically after one correct recall, (b) automatically after two correct recalls, or (c) under participant control (the third condition replicated the previous experiments).

Method

Participants and materials. The participants were 57 UCLA undergraduates who participated for course credit. There were 21, 17, and 19 participants in the User-control, Autodrop-1 and Autodrop-2 conditions, respectively. The materials were the same as the materials in the previous experiments.

Design. The experiment was a 2 (Study-control: Drop vs No-drop) $\times$ 3 (Drop rule: User-control, Autodrop-1, Autodrop-2) mixed design. Items were never dropped in the No-drop condition, which was the same in all three between-participant conditions. The Drop condition differed across groups. In the User-control group, participants were allowed to determine whether or not they dropped items; in the Autodrop-1 condition the computer dropped items automatically after one correct response; in the Autodrop-2 condition, the computer dropped items automatically after two correct responses.

Procedure. The User-control condition was a replication of Experiments 1 and 2, using the trial structure of Experiment 3: On each trial participants were shown a Swahili cue word for 3 seconds, and, except on the first encounter with each pair, they were asked to type in its English translation; then they were shown the correct

[2] One might hypothesise that the small number of items recalled multiple times and then dropped reflected a decision not to drop items recalled multiple times. The opposite was true: Participants were more likely to drop items that they had recalled multiple times (83%) than items recalled less than twice (66%).

[3] Accuracy was computed separately for each participant, and then the participants' scores were averaged. Only 14 of the 22 participants, who had at least one observation at each of the three levels, could be included in the analysis. Items dropped after three or four correct responses could not be included due to a lack of observations.

answer. If the list was assigned to the Drop condition, participants were then allowed to decide whether to continue studying the item, or drop it. However, if the list was assigned to the No-drop condition, participants—unlike in the prior experiments—were not required to press ‘‘Study again later’’ at the end of each trial. That requirement was removed to make the No-drop condition consistent with the two autodrop conditions, in which participants were never shown the ‘‘Study again later’’ or ‘‘Remove from stack’’ buttons.

The Autodrop-1 condition was the same as the User-control condition, except that participants could not choose to drop an item; instead, the program dropped items automatically after they were answered correctly once. In the Autodrop-2 condition, items had to be answered correctly twice, not necessarily consecutively, to be dropped. As in Experiment 3, misspelled answers were considered correct during the study phase and the final test.

Results and discussion

Participants whose items had all been dropped before the end of the list in the Drop condition were excluded from all analyses. The number of participants who were excluded in the User-control, Autodrop-1, and Autodrop-2 conditions, respectively, was 7, 9, and 3, leaving 14, 8, and 16 participants, respectively.

Participants in the User-control condition displayed the same tendency to drop items quickly as did participants in Experiment 3: 68% of the items were dropped after only one correct response, and an additional 8% were dropped without having been answered correctly at all.

Final test accuracy was analysed using a 3 (Drop rule) × 2 (Study-control) ANOVA. As Table 1 shows, there was a significant effect of Drop rule on final test accuracy, $F(2, 35) = 4.44$, $p < .05$, $MSE = .13$, $\eta_p^2 = .20$, with participants in the Autodrop-1 condition showing relatively poor performance. (Note, though, that comparing performance between participants is problematic because different numbers of participants were excluded from the analyses in the different conditions; a problem that, fortunately, does not apply to the within-participant comparison of Drop vs No-drop.) More importantly, there was a significant effect of Study-control: Final test accuracy was higher in the No-drop condition than the Drop condition, $F(1, 35) = 5.87$, $p < .05$, $MSE = .030$, $\eta_p^2 = .14$. Although the interaction was not significant, $F(2, 35) = 1.93$, $p = .16$, $MSE = .030$, $\eta_p^2 = .10$, the Autodrop-1 condition appears to have contributed heavily to the main effect.

TABLE 1
Mean proportion correct (SD) on the final test in Experiment 4 as a function of Study-control and Drop rule

	Drop rule		
Study-control	*User-control*	*Autodrop-1*	*Autodrop-2*
No-drop	.67 (.19)	.44 (.38)	.56 (.32)
Drop	.65 (.24)	.21 (.16)	.50 (.32)

A planned comparison showed that for participants in the Autodrop-1 condition, final test accuracy was significantly higher for lists on which no items were dropped (in the No-drop condition) than it was for lists on which items were dropped after 1 correct response (in the Drop condition), $t(7) = 2.58$, $p < .05$. Final test accuracy was also higher in the No-drop condition than it was in the Drop condition for participants in the Autodrop-2 condition, although the effect was not significant, $t(15) = .85$, $p = .41$. In the User-control condition, final test accuracy was higher in the No-drop condition than it was in the Drop condition, although the difference did not approach significance $t(13) = .35$, $p = .74$.

In summary, the results demonstrated that dropping items after a single correct recall was a maladaptive strategy. Nevertheless, participants in the User-control condition dropped the majority of their items after a single correct recall, replicating Experiment 3. The small but consistent disadvantage of allowing participants to drop flashcards while studying was also replicated, although the difference did not reach statistical significance.

GENERAL DISCUSSION

We found that participants did not profit from being allowed to self-regulate their study time by dropping items. If anything, dropping resulted in a small but consistent disadvantage. The disadvantage was not significant in every analysis, nor was it large in numerical terms, but it is truly surprising because there is a compelling reason to expect the opposite: Dropping ostensibly known items allowed participants to focus more study time on items that they did not know. The average student would find the idea of spending equal time on all information when studying—even information they feel they already know—very foolish indeed. The participants were under no obligation to drop items but did so, presumably, because they believed that doing so would confer an advantage.

The fruitlessness of being allowed to drop items appears to be traceable to poor decision making, not to poor metacognitive monitoring. Participants' relatively good monitoring, as measured by the resolution and calibration of their JOLs in Experiment 2, was coupled with non-optimal decisions as to what items to drop and when to drop them. Other factors may have played a role, such as the reduced spacing of study trials on remaining items as other items are dropped, but the principal implication is that people misunderstand some basic aspects of forgetting and learning, and, therefore, how to manage their study activities.

Types of flawed decision making

Being able to drop items had especially negative effects for the 20% of participants whose study strategy was to drop items they judged difficult to learn. Those participants seem to have believed, mistakenly, that they would not have sufficient time to learn the difficult items. The RPL model suggests that, if there is insufficient time to learn a difficult item, dropping it can be a good decision (e.g., Metcalfe & Kornell, 2003, 2005). The participants' error was that, in reality, they *did* have sufficient time to learn the difficult items. A similar error has been demonstrated when people have been asked to predict how much they will learn by studying once or, for example, four times: Despite large differences in actual learning, the predictions are essentially the same (Kornell & Bjork, 2006). Thus there is one exception to the assertion that participants' metacognitive monitoring was accurate: Some participants seemed to underestimate their ability to learn difficult items across multiple study opportunities.

What about the remaining 80% of participants—that is, those who reported dropping items they knew or found easy? Assuming that dropping items is a potentially useful strategy, why did they fail to profit from being allowed to do so? They, too, may have undervalued the impact of future study opportunities. Perhaps the most surprising result of the current experiments is that participants dropped items that they did not believe they had learned well enough to remember on the final test. The nature of their JOL ratings suggests that their strategy was "I know this now, so I'll drop it, even if I might not get it on the test later." If, as Experiment 3 suggests, such participants did not realise the benefits of continuing to study and test oneself past the point when one can initially produce an answer, it points to their having a fundamental misunderstanding of how learning works. In fact, it is precisely those just-retrievable items, according to the RPL model, that are most learnable, and thus that should *not* be dropped.

Finally, what about the participants who dropped all of the items before the time allocated for studying the list had expired? From one perspective, they should be excluded from the analysis. From another perspective, however, they illustrate some additional perils of self-regulated study, and thus should be included. During the post-experimental debriefing, for example, one participant said that she was well aware—as the instructions made clear—that a blank screen would follow if she dropped everything, but did so anyway because she did not think it would help to study the items any longer. Moreover, students are often motivated—by time and other pressures—to stop studying as soon as possible. In fact, some students drop cards not to allow more time for others, but rather to hasten the end of a study session, because they refuse to stop studying until they have dropped all of their cards.

A final consideration is that because dropping decreases spacing between items, it increases performance levels in the short term (though not necessarily in the long term; see Bjork, 1994, 1999). The increased performance owing to reduced spacing has the potential to increase students' confidence, leading them to stop studying sooner than they otherwise would. Thus, perhaps the fact that some participants spent less time studying when they were allowed to drop cards than when they were not is a realistic feature of the present experiments; one that also argues for including all participants in the analysis.

Practical recommendations

The current findings suggest that the effectiveness of dropping flashcards depends on students becoming metacognitively sophisticated as learners.

Dropping has the potential to be effective, but students need to understand the value of further study, including that—as suggested by the RPL model—items that can be remembered now, but that may be forgotten later, should be given the highest priority, not dropped (Metcalfe & Kornell, 2003). They need to learn, too, the benefits of continuing to be tested on items that one can already recall (see Karpicke & Roediger, 2007).

In the interests of creating durable learning, items, if dropped, should be returned to later. Restudying previously dropped items provides additional spaced-learning opportunities on those items. It also identifies items that have not actually been learned and are in need of further study. It is important for students to realise that items that seem "learned" may be forgotten. Informal conversations reveal that some students return to dropped flashcards and some do not. Perhaps the optimal way of returning to dropped items is via an expanding schedule (Landauer & Bjork, 1978), with increased spacing between each successive study trial. An expanding schedule places less and less emphasis on items that have been studied, allowing more study time on other items.

Students need to understand, too, that a danger of dropping is that it results in a stack of flashcards that has fewer and fewer cards, resulting in decreasing spacing between repetitions of given item and relatively (and often unrealistically) easy recall during study. Such easy retrievals, which are of limited value in terms of fostering long-term recall, can result in illusions of learning. Introducing difficulty, by increasing the number of flashcards in a stack (and the spacing between them), should facilitate long-term learning (Kornell & Bjork, 2007b).

Finally, on the positive side, it is important that students understand that studying with flashcards has important virtues. It incorporates, in a natural way, both testing and spaced practice, two features that, when combined, support both efficient learning and accurate metacognitive monitoring.

Concluding comment

In general, psychologists tend to think of self-regulated study as involving decisions about how and when to study. The present findings demonstrate, however, that an equally important factor in efficient self-regulation of study is deciding when to stop studying—deciding when enough is enough, so to speak (see Kornell & Bjork, 2007a). The results also demonstrate that such decisions require not only complex monitoring and control processes, but also an understanding of how people learn.

Manuscript received 23 January 2007
Manuscript accepted 19 October 2007
First published online 14 November 2007

REFERENCES

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21*, 627–635.

Benjamin, A. S, Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

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.

Christina, R. W., & Bjork, R. A. (1991). Optimising long-term retention and transfer. In D. Druckman & R. A. Bjork (Eds.), *In the mind's eye: Enhancing human performance* (pp. 23–56). Washington, DC: National Academy Press.

Dunlosky, J., & Nelson, T. O. (1992). Importance of kind of cue for judgements of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374–380.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect people's self-paced study and yield "labor-and-gain" effects. *Acta Psychologica*, *98*, 37–56.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, *32*, 779–788.

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

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgements of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 595–608.

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

Kornell, N., & Bjork, R. A. (2006, November). *Predicted and actual learning curves.* Paper presented at the 47th annual meeting of the Psychonomic Society, Houston, TX.

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

Kornell, N., & Bjork, R. A. (2007b, May). *On the illusory benefits of easy learning: Studying small stacks of flashcards.* Poster presented at the 19th annual meeting of the Association for Psychological Science, Washington DC.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 609–622.

Kornell, N., & Son, L. K. (2006, November). *Self-testing: A metacognitive disconnect between memory monitoring and study choice.* Poster presented at the 47th annual meeting of the Psychonomic Society, Houston, TX.

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). London: Academic Press.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

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

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 851–861.

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., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgements in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207–213.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Pyc, M. A., & Rawson, K. A. (in press). Examining the efficiency of schedules of distributed practice. *Memory & Cognition*.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 435–451.

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

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 907–912.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.

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

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