Learners' choices and beliefs about self-testing

Nate Kornell

University of California, Los Angeles

Lisa K. Son

Barnard College

In press at *Memory*. Please do not quote without permission.

Address correspondence to:

Nate Kornell

UCLA Department of Psychology

1285 Franz Hall

Los Angeles, CA 90095-1563

Phone: 310-825-7028

Fax: 310-206-5895

Email: nkornell@ucla.edu

Abstract

Students have to make scores of practical decisions when they study. We investigated the effectiveness of, and beliefs underlying, one such practical decision: the decision to test oneself while studying. Using a flashcards-like procedure, participants studied lists of word pairs. On the second of two study trials, participants either saw the entire pair again (pair mode) or saw the cue and attempted to generate the target (test mode). Participants were either asked to rate the effectiveness of each study mode (Experiment 1) or to choose between the two modes (Experiment 2). The results demonstrated a mismatch between metacognitive beliefs and study choices: Participants (incorrectly) judged that the pair mode resulted in the most learning, but chose the test mode most frequently. A post-experimental questionnaire suggested that self-testing was motivated by a desire to diagnose learning rather than a desire to improve learning.

Keywords: Self-testing, testing effect, flashcards, judgments of learning

Acknowledgments

We thank Robert A. Bjork for his guidance and support and Bridgid Finn for her suggestions and help in conducting the experiments. Grant 29192G from the McDonnell Foundation to Robert A. Bjork supported this research.

Learners' choices and beliefs about self-testing

Self-regulated learning (e.g., homework) requires students to make scores of decisions about how to study. Laboratory research has shown that these decisions are affected by many factors, including the difficulty of the materials (Metcalfe, 2002); the student's goals (Thiede & Dunlosky, 1999); the amount of time pressure the student is under (Metcalfe & Kornell, 2003); a student's age (Kuhn, 2002); and, of course, the type of decision to be made, whether it be strategic (e.g., should I make flashcards?), item-specific (e.g., do I need to review this chapter again?), or global (e.g., should I study tonight?). Many of the practical decisions that students face have not been addressed by experimental research, however (see, e.g., Kornell & Bjork, 2007). One such practical question, which is the focus of the present research, is whether or not to test oneself while studying.

A large body of data suggests that self-testing enhances memory. This so-called *testing effect* refers to the finding that taking a test improves learning more than passively reading the same information (Glover, 1989; McDaniel & Fisher, 1991; Metcalfe, Kornell, & Son, 2007; Roediger & Karpicke, 2006b). The testing effect has been shown to occur when the test is followed by corrective feedback (e.g., Carrier & Pashler, 1992; Cull, 2000), and when it is not (Carpenter & DeLosh, 2005; Landauer & Bjork, 1978). Moreover, testing appears to be especially effective for promoting long-term learning (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006a).

In addition to enhancing memory directly, testing has a second benefit: It allows learners to accurately diagnose what they do and do not know. Making such diagnoses accurately can play an important role in guiding study decisions. Learners' diagnoses of what they know are often assessed via Judgments of Learning (JOLs)—that is, ratings of how probable it is that one will remember an item on a future test. The correlation between the JOLs and actual memory increases when learners can test themselves as they study (e.g., Dunlosky & Nelson, 1992). To the degree that JOLs guide study decisions, accurate JOLs promote effective study decisions, including decisions about which items to study and how long to spend studying (Kornell & Metcalfe, 2006; Nelson, Dunlosky, Graf, & Narens, 1994), and even when to schedule study (Son, 2004).

Beliefs Concerning Self-Testing

Two recent studies suggest that people do not recognize the benefits of testing—instead, they seem convinced that presentations result in more learning than does testing (Roediger & Karpicke, 2006a; Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). In both studies, participants who read a text passage multiple times gave higher JOLs than participants who read the passage and then took one or more tests (Roediger & Karpicke, 2006a, used free recall tests; Agarwal et al., 2008, used cued recall tests). This evidence suggests that, at least in the context of reading text passages, participants thought re-reading was more effective than testing.

There is evidence that in some situations people recognize the benefits of tests. In experiments on the *generation effect*—the finding that generating an answer, for example by unscrambling an anagram or filing in missing letters in a word, leads to better later recall than seeing the word presented whole (Hirshman & Bjork, 1988; Jacoby, 1978; Slamecka & Graf, 1978)—participants have been shown to give higher JOLs for words that they generated than for words that were presented (Begg, Vinski, Frankovich &

Holgate, 1991; Mazzoni & Nelson, 1995)—or to give equal ratings on both types of trials (Maki, Foley, Kajer, & Willert, 1990).

Thus there is evidence to suggest that in some situations, people believe they can learn more via re-reading than being tested. But do students choose to test themselves? To investigate this question, Son (2005) asked first grade students to study a set of cue-target synonym pairs (e.g. *occupation - job*). The participants were asked to judge how well they knew each pair, and then to decide, for each pair, whether to re-read the pair or test themselves. The first graders chose to test themselves on items that they felt they knew well, and preferred re-presentation on less well-known items, although they may have chosen tests on items that they thought could answer correctly, not as a study strategy but simply to impress their teachers.

Choosing Self-Testing: A Pilot Experiment

We wanted to follow up on the question of self-testing using a procedure that seemed to more strongly resemble real-world learning. Thus, we used a flashcard-like procedure. When learners use flashcards, which are a ubiquitous study tool (Kornell & Bjork, 2007; Kornell, in press), they generally cycle through a set of cards repeatedly, looking at the front of each card, testing themselves, and then looking at the back. Flashcard-like programs are becoming increasingly popular online. Asking participants to study computer-based flashcards seemed to be an appropriate experimental context in which to investigate self-testing.

In the pilot study, the learning materials were English-Indonesian vocabulary (e.g., *Left-Kiri, Hot-Panas, Late-Terlambat*), which were presented for study on a computer. Participants were allowed to decide how they studied: They could choose to read the pairs intact (*pair* mode), or to have the cue presented alone before the target appeared (*test* mode). If a participant changed modes, which they could do at any time using buttons on the computer screen, the remaining items in the list would be presented in whichever mode was selected, unless the mode was changed again. There were three between-participant conditions: In addition to the condition in which participants could choose the study mode, there was an all-pair mode, in which there were no test trials, and an all-test mode, in which there were no pair trials. In the same way that students doing homework are under time pressure, our participants were given a time limit of 10 minutes to study, and they controlled the timing of presentations as they studied. The fixed time limit gave participants an incentive to manage their time wisely, because studying quickly translated into more study trials later (unlike most previous self-regulated study research; see Son & Metcalfe, 2000). There was a final test at the end of experiment.

The results from the pilot study showed that when given the choice, participants began by studying in presentation mode but quickly switched to test mode. There was also a testing effect: On the final test, the recall accuracy of participants in the choice condition—who only tested themselves on an average of only 55% of the trials—matched that of participants in the all-test mode (mean performance = 63%), and surpassed that of participants in the all-pair mode (mean performance = 37%). Thus it appears that a mix of presentation trials and test trials were no less effective than constant test trials.

Participants in the pilot experiment chose to test themselves The question addressed in the present research was why did they choose to do so? Tests enhance learning, as the pilot experiment demonstrated, but another reason to test oneself is as a way to *diagnose* one's memory. In a survey of 472 UCLA undergraduates (Kornell &

Bjork, 2007), for example, 91% reported testing themselves regularly while studying. Out of all respondents, 68% reported that they test themselves primarily to make a diagnosis of what they do and do not know. Thus students seem to be aware of the metacognitive benefits of testing (also see Dunlosky, Serra, Matvey, & Rawson, 2005). They seem less aware of the direct memory benefits of testing, however: only 18% reported testing themselves to enhance their learning directly. Thus the majority of students seem to view tests the way their teachers might: primarily as assessments, not learning tools.

The Present Experiments

In sum, previous research provides a puzzling picture of learners' attitudes towards testing: People believe testing is less effective than re-reading (e.g., Roediger & Karpicke, 2006a), and yet, as our pilot results demonstrate, they choose self-testing preferentially over re-presentation nonetheless. One explanation of these seemingly contradictory findings is that people choose testing to diagnose their learning, not to enhance it (Kornell & Bjork, 2007). Another explanation, however, is a difference in the learning materials and test-type: People rated re-reading as effective when studying passages (Roediger & Karpicke) but chose self-testing while studying translations on digital flashcards (in the pilot study). Study strategies vary widely depending on the subject matter and expected type of test—to take an obvious example, literature students read more and solve problems less than physics students—and how people rate testing as in one context cannot necessarily be used to predict how they will rate, or utilize, testing in another context.

In the present experiments we investigated choices and beliefs regarding self-testing. Two experiments, which shared a single set of materials, differed in only one aspect of their procedures. In Experiment 1, participants were asked to judge the relative effectiveness of testing versus presentation. In Experiment 2, participants were asked to choose between testing and presentation. Based on previous research, we predicted that participants would choose self-testing rather than presentation. In the same situation, however, we also predicted that participants would rate presentation as benefiting their learning more than testing— which would demonstrate a mismatch between study choices and metacognitive judgments (JOLs).

Feedback was also manipulated in the present experiments. Feedback plays an important role in the benefits of tests (e.g. Metcalfe & Kornell, 2007; Pashler Cepeda, Wixted, & Rohrer, 2005): Without feedback, if one is unable to answer a question initially, there is little hope of recovering it later, whereas when feedback is provided, errors can be corrected, and tests are endowed with the benefits of a presentation while also maintaining the additional benefits of testing. Thus we predicted that participants would choose testing more often, and rate testing more favorably, when feedback was provided than when it was absent.

Experiment 1

In Experiment 1, participants studied a list of word pairs twice. The first study trial was always in pair mode (i.e., the cue and target were presented together). In the *pair* condition, the second study trial was also in pair mode; in the *test* condition, the second study trial was in test mode (i.e., the cue was presented and participants were asked to type in the target). After the second trial, participants were asked to predict how many of the pairs they would remember on a final test that would occur a short time later. This aggregate JOL allowed us to examine participant's' metacognitive beliefs about the relative effectiveness of testing versus presentation. We also manipulated whether feedback was given during the test.

The hypothesis, in Experiment 1, was that participants would rate the pair condition as more effective than the test condition. Such a finding would replicate previous findings (e.g., Roedger & Karpicke, 2006a), but it would also differ from previous work in important ways. Firstly, Roediger and Karpicke's task, free-recalling text passages without feedback—which students rarely do when studying—is probably less attractive than testing oneself in a flashcards-like paradigm, in which students naturally do (Kornell & Bjork, 2008). Moreover, in Roedger and Karpicke's study, the participants' predictions actually accurate with respect to the immediate test—that is, consistent with the participants' predictions, on an immediate test, study trials resulted in better performance than test trials. The predictions were inaccurate with respect to Roediger and Karpicke's delayed test condition, but people often fail to predict that forgetting will occur at all (Koriat, Bjork, Sheffer, & Bar 2004), much less differential forgetting rates following test trials compared to presentation trials. Thus, in addition to providing a point of comparison for Experiment 2, Experiment 1 differed in important ways from previous results regarding people beliefs about the benefits of testing.

Method

Participants. Thirty-five college students participated for course credit, 19 in the feedback condition and 16 in the no-feedback condition.

Materials. The materials were 48 English word pairs. Half were relatively easy, related pairs (e.g., whale-mammal), which had forward association strengths of between .05 and .054 based on free association norms (Nelson, McEvoy, & Schreiber, 1998). The other half were more difficult, unrelated pairs (e.g., inanity-capacity), with zero forward association strength.

Design. We used a 2x2 mixed design. Mode (pair or test) was manipulated within participants; Feedback (feedback vs. no feedback) was manipulated between subjects.

Procedure. The procedure comprised three phases: study, distractor, and test. There were 4 lists of 12 word pairs each; participants studied and made judgments about each list individually, and then, after the distractor task, took a test on all four lists. Two of the lists were assigned to the all-pair mode (either the first two lists or the last two lists) and two were assigned to the all-test mode. Within each list, half of the items were easy and half were difficult.

The study phase for a given list comprised three phases: initial presentation, restudy, and JOL. During initial presentation, the 12 word pairs were presented for study, in random order, for 5 seconds each. During restudy, for lists presented in the pair condition, the full word pairs were presented again, for 5 second each; for lists presented in the test condition, the cue word was presented alone, and participants were asked to type in the target word. Feedback was provided (i.e., the correct answer was presented for 1 second after the participant responded) to participants in the feedback condition. The JOL phase followed restudy. Participants were asked to make an aggregate JOL (i.e. a judgment about an entire list, rather than a single item) by completing this sentence: "When I am tested on that word list in about 15 minutes, I think I will get about __ out of 12 correct."

After the fourth list was presented, there was a 5-minute distractor task, during which participants were asked to identify well-known people based on photographs presented upside-down.

The test followed the distractor task. Each cue was presented, one at a time, and participants were asked to type in the answer and press return. The items from list one were tested first, in random order, followed by the items from list two, and so on.

After completing the test, participants were asked "Which best describes the reason you quiz yourself when you study?" There were four response options: a) I quiz myself because I learn more that way than I would through presentation, b) I quiz myself to figure out how well I have learned the information I'm studying, c) I quiz myself because I find quizzing more enjoyable than presentation, d) None of the above.

Results and Discussion

Final test accuracy was higher for items studied in the test mode than the pair mode, $F(1, 33) = 7.60$, $p < .01$, $\eta_p^2 = .19$ (Figure 1a). The effect of feedback was not significant ($F(1, 33) = .60$, $p = .44$), nor was the mode x feedback interaction ($F(1, 33) = .40$, $p = .53$). In opposition to actual accuracy, JOL ratings were higher for items in the pair mode than the test mode, $F(1, 33) = 7.74$, $p < .01$, $\eta_p^2 = .19$ (Figure 1b). Again, the effect of feedback was not significant ($F(1, 33) = 1.13$, $p = .30$), nor was the interaction ($F(1, 33) = .34$, $p = .57$). Thus testing enhanced learning, but participants rated extra presentations as more effective than tests. In addition, the survey showed that the main reason participants chose to self-test was to diagnose their level of learning, not to improve it (see Table 1). When interpreting the post-experimental question, it is important to remember that participants were allowed to select only one response; thus

Table 1 presents participants' main, but not necessarily only, reason for testing themselves.

Experiment 2

Experiment 2 was identical to Experiment 1, with one exception. In Experiment 1, lists were randomly assigned to either the pair mode or test mode. In Experiment 2, the participants were allowed to choose between the two modes (as they had done in the pilot experiment). During the study phase, after studying a list for the first time, participants were asked "Now it's time to study that list again. What would you like to do: see the pairs presented again, or take a practice quiz?" They selected one of two buttons, corresponding to the pair condition and the test condition respectively, labeled "Present again" and "Practice quiz." Thirty-five college-aged students participated for course credit, 16 in the feedback condition and 19 in the no-feedback condition. It should be noted that the two conditions transpired during the summer and fall semesters, respectively, and thus participants could not be assigned to conditions randomly.

Results

As Figure 1c illustrates, on the first list, participants chose the pair mode and test mode equally as often, each at a rate of 50% (and there was no difference between the feedback and no-feedback groups). With experience, however, participants developed a preference for testing. There was a significant effect of list ($F(1, 33) = 4.43$, $p < .01$, $\eta_p^2 = .12$), but no significant effect of feedback overall ($F(1, 33) = .64$, $p = .43$), and no feedback x list interaction ($F(1, 33) = .85$, $p = .47$).

To examine participants' preferences for the test mode versus the pair mode statistically, we chose to examine study choices on list 4, because that is the list on which participants had the most experience with the experimental procedure, and thus could make the most informed choices. A planned comparison showed that participants in the feedback condition chose testing at a rate significantly above 50%, $t(15) = 4.39$, $p < .001$, $d = 1.10$. The effect was not significant in the no-feedback condition, $t(18) = 1.16$, $p = .26$. Choosing testing at an especially high rate when tests were followed by feedback seems adaptive given that in the absence of feedback, participants have little chance to learn items that they cannot recall. However, comparing the feedback condition and the no-feedback condition, there was not a significant difference in the rate at which participants chose testing, $t(33) = 1.66$, $p = .11$. When the data from the two feedback conditions were collapsed, the data demonstrated a significant preference for testing, $t(34) = 3.24$, $p < .01$, $d = .55$.

Final test accuracy and JOLs were not the focus of Experiment 2, although we report the data for completeness. The findings should be interpreted with caution for two reasons, however: First, because the participants assigned themselves to conditions, and second, because 16 of the 35 participants were excluded from the analyses, 11 because they chose the test mode exclusively and 5 because they chose the pair mode exclusively. These concerns do not apply to Experiment 1, which is why Experiment 1 was conducted, and why it serves as the basis for the claim that participants rated presentation as more effective than testing. Nevertheless, the results of Experiment 2 were consistent with the results of Experiment 1: Participants gave higher JOLs following the pair mode than the test mode ($F(1, 17) = 16.40$, $p < .001$, $\eta_p^2 = .49$), despite performing in the opposite manner. Final test accuracy was numerically higher for items in the test mode than in the pair mode, but the difference was not significant, $F(1, 17) = 1.34$, $p = .26$, $\eta_p^2 = .07$. Feedback did not have significant effects on JOLs or final test accuracy.

A final analysis combined Experiments 1 and 2 in examining participants' responses to the post-experimental survey. As Table 1 shows, the majority of participants reported testing themselves to diagnose their memories, not to enhance them directly, consistent with the findings of Kornell and Bjork (2007).

Discussion

Participants in Experiment 2 chose to test themselves rather than to receive re-presentations as they studied. This finding demonstrates a mismatch between metacognitive beliefs and study choices: Participants in Experiment 1 believed, wrongly, that testing was less effective than presentation, but in the same situation, participants in Experiment 2 chose testing rather than presentation. Survey results suggest that the reason people chose to test themselves, despite believing that doing so impaired their learning, was as a way to diagnose their own learning.

General Discussion

The current experiments examined learners' use of self-testing as a study strategy. Taken together, Experiments 1 and 2 showed that, in a paradigm that resembled studying computer-based flashcards, participants preferred testing themselves rather than re-presentation—and benefited from doing so—but judged re-presentation as more effective than self-testing. Thus there was a mismatch between study choices, which favored testing, and metacognitive beliefs, which favored presentation. The data also suggest that people test themselves in order to diagnose what they do and do not know, often without realizing that doing so enhances their memory.

The Benefits of Tests

There has been a recent resurgence in research on the testing effect (e.g., Roediger & Karpicke, 2006b). When different degrees of testing have been compared, the maximum amount of testing has generally resulted in the maximum benefit (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006a; but see also Izawa 1970). The current results, however, indicate that more testing may not always be better: The pilot data showed that being tested 55% of the time resulted in performance that was as good as being tested 100% of the time. This finding implies instead that *when* a test occurs may be as crucial as the mere fact *that* it occurs.

Why did high levels of learning result when participants chose to test themselves on only 55% of the trials, primarily at the end of the study period? One advantage of beginning in presentation mode may have been that presentations communicate new information more quickly, and efficiently, than a test, because tests require two steps, a question followed by feedback, whereas a presentation requires only one step (Izawa, 1992). Moreover, early in the study phase unknown items were predominant, and for such items presentations and tests may be similarly valuable—although retrieval attempts seem to enhance learning, relative to presentations, even for unknown items (see Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009)—because tests can take more time to complete than presentations. As the study phase went on, however, the participants learned an increasing number of the items. When an item can already be recalled, additional presentations, at least in some circumstances, appear to confer little or no benefit, whereas additional tests can have large effects (Karpicke & Roediger, 2007, 2008). Thus as time went on, and the participants became able to recall more items, choosing the presentation mode would have become increasingly unwise. In short, participants may have benefited from presentations because of efficiency initially, and then benefited from the mnemonic benefits of tests, relative to the ineffectiveness of re-studying, later in the study phase when they knew many of the items.

Although the participants' choosing to self-testing may have enhanced their learning, that is not necessarily why they chose to test themselves, as Experiments 1 and 2 demonstrate. It is certainly possible that participants were astutely attempting to optimize their learning. It is more likely, based on the post-experimental questions in Experiments 1 and 2, that participants chose testing to discern what they did and did not know. A third possibility is that participants simply wanted to answer correctly, and to serve that goal, they waited to test themselves until they felt that they would be able to do well on the tests.

A Metacognitive Dissociation

Research on memory monitoring is often motivated by the idea that memory monitoring guides study decisions (Benjamin, Bjork, & Schwartz, l998; Dunlosky & Hertzog, 1997; Dunlosky & Thiede, 1998; Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2003, 2005; Son, 2004; Son & Metcalfe, 2000). A number of studies have established that study decisions are based on JOLs (e.g., Thiede & Dunlosky, 1999; Son & Metcalfe, 2000; although Koriat, Ma'ayan, & Nussinson, 2006, suggest that JOLs are based on study decisions). Our participants, by contrast, chose a study technique (testing) that they thought was less effective than the alternative (presentation). This finding suggests that memory monitoring was not the primary basis for the study decisions that were made in the current experiments.

In conjunction with prior research, the current data suggest that there is no single answer to the question of what guides study decisions. A satisfactory theory of study-time allocation should reflect the fact that students' goals include more than just finding the study strategy that will have the largest direct benefit for learning (Flavell, 1979; Thiede & Dunlosky, 1999). Our participants tested themselves to diagnose their memories—a worthy goal, especially because doing so can guide future study decisions. Ironically, though, our participants' decisions were at odds with the standard assumption—that is, that people study with the goal of maximizing their learning directly. The difference between the current findings and previous experiments may be related to the fact that we asked people to make a strategic decision about how to study (i.e., whether or not to test themselves), as opposed to item-based decisions about whether and for how long to study a specific word pair.

There is precedent for the mismatch between memory monitoring and study choices. For example, Moulin, Perfect, and Jones (2000) found a dissociation between judgments of learning and study time allocation in participants with Alzheimer's disease. Lee (2005) also showed such a dissociation; participants reported that they went to lecture before reading their textbooks, despite thinking that reading the textbook and then going to class was more effective—probably because they also rated reading the textbook first as more difficult than going to lecture first.

Conclusion

Researchers often assume that people try to study in ways that maximize learning. In most situations, that assumption is surely valid. Participants in the present experiments chose a sub-optimal study strategy, however—they chose to test themselves, despite believing that re-presentation would do more to improve their memories. Learners study in suboptimal ways in a variety of situations—for example, by prematurely ceasing to study information that they do not yet know (Kornell & Bjork, 2007)—but in most cases, the learners are trying to study optimally, but they do not understand what strategy would be most effective. Such a misunderstanding occurred in the present results; the learners thought presentation was more effective than testing. The unique aspect of the current findings is that the strategy learners chose was the one that *they believed* was least effective. They did not try to maximize learning; instead they tried to maximize the accuracy of their metacognitive monitoring. By demonstrating that learners sometimes prioritize enhancing their ability to monitor their learning over enhancing their learning itself, the current results underscore the importance of goals in determining how people study.

References

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

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition, 19*, 487-497.

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. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical Aspects of Memory: Current Research and Issues* (Vol. 1, pp. 396-401). New York: Wiley.

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215-235.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *Journal of Gerontology: Series B: Psychological Sciences & Social Sciences, 52B*, 178-186.

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

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *Journal of General Psychology, 132*, 335-346.

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.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Glover, J. A. (1989). The ''testing'' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399.

Hirshman, E. L., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*, 484-494.

Hogan, R.M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340-344.

Izawa, C. (1992). Test trials contributions to optimization of learning processes: Study/test trials interactions. In A. F. Healy & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes: Vol. 1. From learning theory to connectionist theory* (pp. 1-33). Hillsdale, NJ: Erlbaum.

Jacoby, L.L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

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.

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

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The Intricate Relationships Between Monitoring and Control in Metacognition: Lessons for the Cause-and-Effect Relation Between Subjective Experience and Behavior. *Journal of Experimental Psychology: General, 135*, 36-69.

Kornell, N. (in press). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*.

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). *Unsuccessful retrieval attempts enhance subsequent learning*. Manuscript submitted for publication.

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

Kuhn, D. (2002). Metacognitive development. *Current Directions in Psychological Science, 9*, 178-181.

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.

Lee, B. G. (2005). Lecture first or text first? Optimizing undergraduate instruction. *Dissertation Abstracts International, 66* (09), 5117B. (UMI No. 3190464)

Maki, R. H., Foley, J. M., Kajer, W. J., Thompson, R. C, & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 609-616.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1263-1274.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*, 192-201.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349 –363.

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., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14*, 225-229.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based program to enhance study efficacy in a high and low-risk setting. *European Journal of Cognitive Psychology, 19*, 743-768.

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000). The effects of repetition on allocation of study time and judgments of learning in Alzheimer's disease. *Neuropsychologia, 38*, 748-756.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

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

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3-8.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). *The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning?* Manuscript submitted for publication.

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

Roediger, H. L., & Karpicke, J. D. (2006b). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science, 1,* 181-210.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory, 4*, 592-604.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 601-604.

Son, L. K. (2005). Metacognitive control: Children's short-term versus long-term study strategies. *Journal of General Psychology, 132*, 347-363.

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.

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.

Author Note

Nate Kornell, Department of Psychology, University of California, Los Angeles, and Lisa K. Son, Department of Psychology, Barnard College.

Correspondence should be addressed to Nate Kornell, at: Department of Psychology, 1285 Franz Hall, UCLA, Los Angeles, CA, 90095. Email: nkornell@ucla.edu

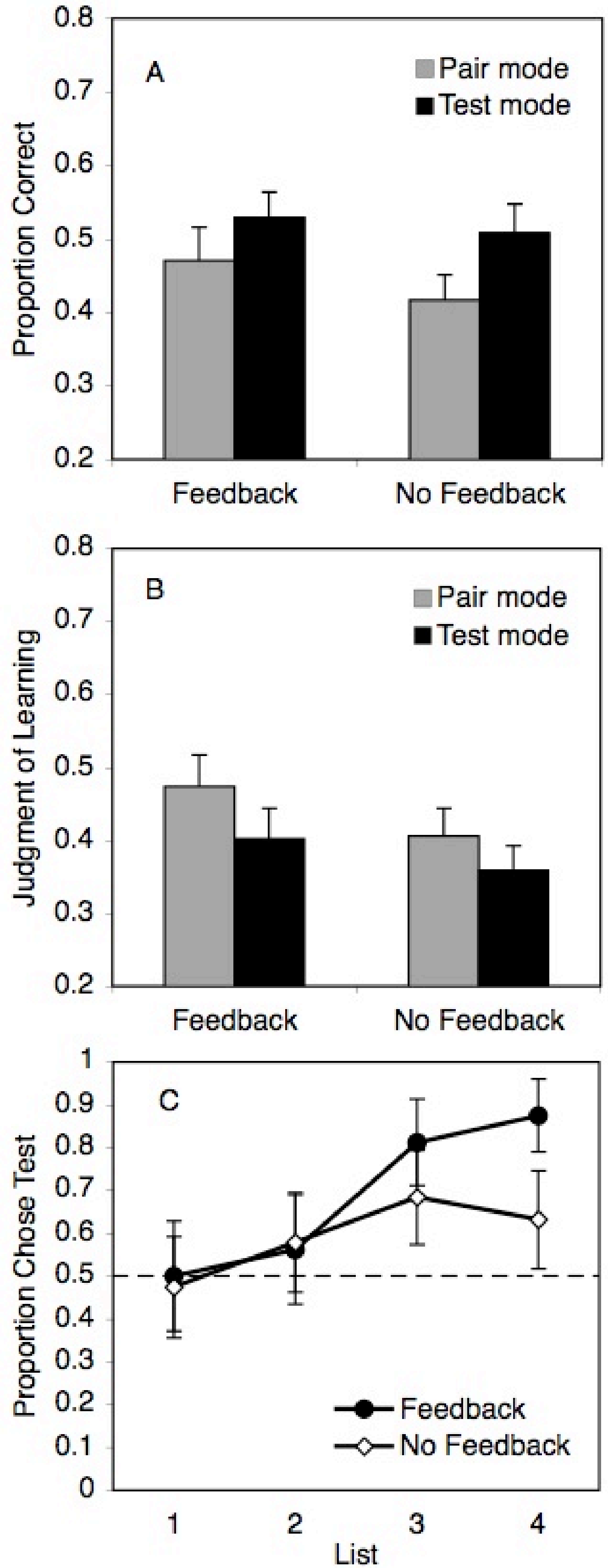
Table 1

Responses to the post-experimental question "Which best describes the reason you quiz yourself when you study?" combined across Experiments 1 and 2.

Multiple choice response	Respondents
I learn more that way than I would through presentation	20%
To figure out how well I have learned the information I'm studying	66%
I find quizzing more enjoyable than presentation	4%
None of the above	10%

Figure Caption

Figure 1. Results of Experiments 1 and 2. A) Proportion correct in Experiment 1. B) Judgments of Learning in Experiment 1. C) Proportion of lists on which participants chose to be tested, as function of list, in Experiment 2 (the dashed line represents indifference to testing vs. presentation). Error bars represent standard errors.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
Proportion Correct
A
Pair mode
Test mode
Feedback
No Feedback
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Judgment of Learning
B
Pair mode
Test mode
Feedback
No Feedback
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion Chose Test
C
Feedback
No Feedback
1
2
3
4
List