

# Metacognitive Control and Strategy Selection: Deciding to Practice Retrieval During Learning

Jeffrey D. Karpicke  
Purdue University

Retrieval practice is a potent technique for enhancing learning, but how often do students practice retrieval when they regulate their own learning? In 4 experiments the subjects learned foreign-language items across multiple study and test periods. When items were assigned to be repeatedly tested, repeatedly studied, or removed after they were recalled, repeated retrieval produced powerful effects on learning and retention. However, when subjects were given control over their own learning and could choose to test, study, or remove items, many subjects chose to remove items rather than practice retrieval, leading to poor retention. In addition, when tests were inserted in the learning phase, attempting retrieval improved learning by enhancing subsequent encoding during study. But when students were given control over their learning they did not attempt retrieval as early or as often as they should to promote the best learning. The experiments identify a compelling metacognitive illusion that occurs during self-regulated learning: Once students can recall an item they tend to believe they have “learned” it. This leads students to terminate practice rather than practice retrieval, a strategy choice that ultimately results in poor retention.

*Keywords:* retrieval practice, metacognition, self-regulated learning, study-time allocation, learning strategies

Research on human learning and memory has traditionally implemented strict experimental control over the mnemonic activities of learners by providing them with an encoding strategy or by inducing a type of processing and examining the effects of such manipulations on learning. A classic example is the work on levels of processing by Craik and Tulving (1975). They carried out 10 experiments in which they exercised precise control over the processing that learners engaged in when studying lists of words. Most research on the effects of organization, elaboration, mental imagery, spaced practice, and other mnemonic strategies has followed this tradition by examining the conditions and factors that promote learning.

Considerably less research has examined the circumstances in which people select effective (or ineffective) learning strategies or the monitoring processes that guide strategy choices (though exceptions exist, e.g., in research on child development; see Pressley & Harris, 2006). Self-initiated strategies are often viewed as a nuisance to be avoided or neutralized because giving subjects control over aspects of their learning conflicts with the desire of experimenters to maintain experimental control over subjects’

behavior (Nelson & Narens, 1994). This state of affairs exists in memory research despite the fact that people must recruit strategies on their own when learning new material in real-world contexts. Although instructors might provide students with strategies in educational settings, surveys of college students indicate that this is rare (Kornell & Bjork, 2007), and because students often receive little or no direct instruction about how to learn they instead develop their own idiosyncratic strategies, which are likely to vary in effectiveness. In short, research asking the question “What strategies promote good learning?” has significantly outstripped research that asks “Do people choose effective strategies when they monitor and regulate their own learning?”

The present research investigated the interrelationships among three factors: (1) the effects of retrieval practice on learning, (2) students’ monitoring of their own learning, and (3) students’ decisions to practice retrieval in self-controlled learning conditions. We know that practicing retrieval is an effective way to improve learning and long-term retention (for review, see Roediger & Karpicke, 2006a). But are students aware of this? Do they choose to practice retrieval when they control their own learning? And how does metacognitive monitoring guide the strategies students choose? Despite considerable research on metacognitive regulation of learning and despite the potent mnemonic effects of testing, self-testing has never been thoroughly investigated (but see Kornell & Son, 2009). This research represents a systematic evaluation of metacognitive monitoring during learning and metacognitive control processes involved in deciding to practice retrieval.

## Retrieval Practice and Learning

Students and educators often view the act of taking a test as a neutral assessment of learning that occurred in a prior study period

---

This article is based on a PhD dissertation submitted to Washington University in St. Louis, Missouri. The research was supported by a Dissertation Research Award from the American Psychological Association and a graduate research scholarship from the American Psychological Foundation/Council of Graduate Departments of Psychology. I thank Roddy Roediger for his invaluable guidance and Mark McDaniel, Kathleen McDermott, Steve Peterson, Keith Sawyer, and Jim Wertsch for serving on my committee.

Correspondence concerning this article should be addressed to Jeffrey D. Karpicke, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081. E-mail: karpicke@purdue.edu

(Roediger & Karpicke, 2006a). Tests are assumed to measure the contents of memory but not to change them. In contrast to this broad assumption, research examining the testing effect has shown that tests not only assess learning but also can enhance long-term retention. The testing effect is not due to the fact that a test constitutes another exposure to the material and the effect occurs even when subjects are not given feedback about their performance (e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006b). The fact that testing without feedback promotes learning indicates that the act of practicing retrieval is itself a mnemonic enhancer (see Bjork, 1975, 1988; Carpenter & DeLosh, 2006; Chan, McDermott, & Roediger, 2006; Karpicke & Roediger, 2007, 2008; Pyc & Rawson, 2009; Roediger & Karpicke, 2006b).

Retrieval can also promote learning by improving encoding during a subsequent study period. This phenomenon was called the *potentiating* effect of testing by Izawa (1969, 1971) and refers to the finding that inserting tests within a learning phase facilitates performance more than repeated studying without testing. In other words, when students attempt to recall items they have not yet learned, the act of attempting retrieval enhances their ability to encode the items during subsequent study. There have been few (if any) investigations of the potentiating effect of testing in the 40 years since Izawa's program of research (see Karpicke & Roediger, 2007; Kornell, Hays, & Bjork, 2009).

Retrieval practice is a potent way to promote long-term retention and encoding during restudy, but do students practice retrieval when they regulate their own learning? What factors would lead students to test themselves, and what might trick students into not practicing retrieval? Given that testing is a powerful way to enhance learning, why would students not practice retrieval while learning?

### Metacognitive Monitoring During Learning

Questions about how people regulate their own learning are often examined within the monitoring and control framework in metacognition, which specifies a causal relationship between metacognitive monitoring and metacognitive control (see Koriat, 2007; Koriat & Goldsmith, 1996; Nelson & Narens, 1990, 1994). Subjects monitor their progress while learning, and their subjective assessments of their learning are usually measured as judgments of learning (JOLs).

Prior research suggests that subjects' JOLs are not sensitive to the mnemonic effects of retrieval practice (see Karpicke & Roediger, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006b). It is therefore important to examine why this might be the case—specifically, what information do subjects use to guide their JOLs? Current theories generally agree that JOLs may be based on a variety of factors or cues, as in Koriat's (1997) cue-utilization framework (see too Dunlosky & Matvey, 2001). There are several potential cues for JOLs, including the perceived ease or difficulty of individual items (called intrinsic difficulty in Koriat's scheme), the number of study/test trials it takes to learn an item (cf. Koriat, 2008), and the processing fluency during a learning task (e.g., encoding and retrieval fluency; see Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005). It is likely that a combination of these factors influences subjects' JOLs and thereby influences their strategy choices.

### Metacognitive Control of Learning

If subjects' JOLs are not sensitive to the fact that retrieval practice produces learning, then when subjects are given control over their own learning, they may not practice retrieval as often as they should. The monitoring and control framework proposes that subjects execute control processes that are guided by monitoring (JOLs). The control process examined most extensively in metacognition research has been study-time allocation (Dunlosky & Thiede, 1998; Metcalfe & Kornell, 2003), which is usually defined as either self-paced study times for individual items or the selection of items to be studied again in a second study period (see Metcalfe, 2009; Son & Kornell, 2008). The present research took a different approach to examining the strategies that subjects select when they regulate their own learning. First, the effectiveness of retrieval practice was established in experimenter-controlled conditions. Then in learner-controlled conditions subjects could choose different strategies—specifically they could choose to practice retrieval, to restudy items, or to remove them from further practice. This experimental approach makes it possible to compare learning performance with subject-selected choices to learning performance under experimentally controlled conditions. Importantly, if subjects' JOLs are not sensitive to the effects of retrieval practice, then subjects likely will not choose to practice retrieval when given control over their own learning.

### Introduction to the Experiments

In the present experiments, subjects learned foreign-language vocabulary word pairs across a series of study and test periods. Subjects made JOLs during the task to assess their monitoring of their own learning. In some experiments, the learning conditions were experimenter-controlled—whether items were repeatedly tested or repeatedly studied or removed was experimentally manipulated. In other experiments the subjects were given control over their own learning and could choose to test (practice retrieval), restudy, or remove particular items.

Experiments 1 and 2 examined JOLs and strategy choices in test periods. In Experiment 1, the learning conditions were experimentally controlled. Once a pair was correctly recalled for the first time it was either repeatedly tested, or repeatedly studied, or dropped from further practice. The prediction was that repeated retrieval practice in the test condition would enhance long-term retention assessed on a final test 1 week after learning (Karpicke & Roediger, 2007, 2008). In Experiment 2, subjects chose strategies for individual items in the task. Once subjects correctly recalled a pair for the first time they then decided whether to repeatedly test, restudy, or drop the pair. The conditions in Experiment 2 were the same as those used in Experiment 1. Thus Experiment 2 examined the metacognitive control policy adopted by subjects—the relationship between JOLs and the particular strategy they chose (cf. Koriat & Goldsmith, 1996). The prediction was that subjects' choices would be tied to their JOLs and that consequently they would not often choose to practice retrieval. This would in turn have negative consequences for long-term retention.

Experiments 3 and 4 examined JOLs and strategy choices in study periods. In Experiment 3 the subjects learned the list of vocabulary words in repeated study periods—without testing after each study period—or in alternating study and test periods. The

prediction was that inserting tests in the learning phase in the alternating study/test condition would produce better learning performance relative to repeated studying without testing. This prediction is based on the idea that attempting retrieval potentiates learning by enhancing encoding during subsequent study (cf. Izawa, 1969). In Experiment 4 the subjects made strategy choices (to test or restudy or remove pairs) during study periods. The question in this experiment was whether subjects would choose to test themselves early—even if they did not anticipate correct recall—because doing so would show that subjects realized that attempting recall would enhance encoding during subsequent study. The prediction was that subjects would not do this and that this choice would have negative consequences for initial learning.

### Experiment 1: Retrieval Practice and Monitoring During Learning

The purposes of Experiment 1 were (a) to gather benchmark data on the effect of retrieval practice on long-term retention and (b) to examine JOLs to see whether subjects had metacognitive awareness of the fact that retrieval practice would enhance retention. Three learning conditions were examined. In the Test condition, once a pair was correctly recalled it was tested two additional times in the next two test periods but no longer studied in subsequent study periods. In the Study condition, when a pair was correctly recalled it was studied again in the next two study periods but not tested in any subsequent test periods. In the Drop condition, when a pair was correctly recalled it was removed from further study and test periods. Subjects made JOLs during test periods, and the analyses focused on JOLs after the first successful recall of each pair. If subjects are aware of the positive effects of retrieval practice then their JOLs should reflect this awareness. Alternatively subjects might believe that once they can recall an item they had “learned” it and their JOLs might not reflect the fact that additional retrieval practice would help promote long-term retention.

#### Method

**Subjects.** Thirty Washington University undergraduates participated in Experiment 1 in exchange for payment or course credit. Three learning conditions were examined (the Drop, Study, and Test conditions), and 10 subjects were assigned to each condition.

**Materials.** A total of 60 Swahili–English word pairs were selected from the norms of Nelson and Dunlosky (1994) for use in the series of four experiments. Normative ratings of ease of learning (EOL) of the pairs were obtained from 25 subjects who did not participate in any of the experiments reported here. The average EOL rating was treated as a measure of the perceived intrinsic difficulty of the items following previous research (Koriat, 1997; see too Leonesio & Nelson, 1990; Nelson & Leonesio, 1988). Subjects in the normative study were asked to rate each of the 60 pairs in terms of how easy or difficult it would be to learn. The subjects made their ratings on a scale from 0 to 10, where 0 indicated the most difficult pair and 10 indicated the easiest pair. The average EOL rating was calculated for each pair, and the intraclass correlation of the average ratings was .84, indicating high interrater reliability (Shrout & Fleiss, 1979).

Forty pairs were used in Experiment 1. The EOL ratings for the pairs averaged 5.1 (median = 5.1,  $SD = 1.3$ ) and ranged from 2.6 to 7.8. On the basis of a median split on the EOL ratings, 20 were deemed easy pairs ( $M = 6.2$ ) and 20 were deemed difficult ( $M = 4.1$ ).

**Procedure.** Subjects were tested in small groups of 5 or fewer. At the beginning of the learning phase they were told they would learn a list of 40 Swahili vocabulary words and their English translations during a series of alternating study and test periods. Each study period consisted of several study trials, and an individual pair was presented in each trial. During study trials, subjects saw a Swahili word with its English translation below it (e.g., *mashua*–boat) on a computer screen and were told to study the pair so that they could recall the English word given the Swahili word. The subjects were told to press the space bar to move on to the next pair in the study period and were instructed to spend as little total time as possible studying the entire list—a standard instruction in self-paced study experiments (cf. Koriat, Ma’ayan, & Nussinson, 2006) intended to prevent subjects from spending extraordinarily long times studying. After every study period the subjects performed a 30-s distractor task that involved verifying multiplication problems to eliminate primary memory effects (Glanzer & Cunitz, 1966).

Each test period consisted of several test trials, and an individual pair was tested in each trial. During test trials, subjects saw a Swahili word with a cursor below it and were told to type the correct English word for each Swahili word (e.g., subjects would be shown *mashua* to recall *boat*). Each test trial lasted 8 s (with a 500-ms intertrial interval), after which the computer program automatically advanced to the next trial regardless of whether the subject had entered a response. Recall time was assessed as the duration between the onset of the cue and the first key press of the subject’s response. Following each test trial the subjects were shown the cue and the response they had entered and were asked to assess how well they had learned the pair by typing a number between 0% and 100%, where 0% indicated they had not learned the pair well and did not believe they would be able to recall it in the future, 100% indicated they had learned the pair extremely well and were perfectly confident they would recall it in the future, and intermediate levels reflected intermediate judgments of learning.

Following the test period, subjects studied the list again in another study period, then were tested on the list again in another test period, and so on until they had correctly recalled all 40 pairs at least one time during the learning phase. The order of trials within each period was randomized by the computer.

Subjects in the three conditions were treated identically up to the point when they correctly recalled a pair for the first time. In the Drop condition, once a pair was correctly recalled it was removed from further study and test periods. In the Study condition, once a pair was correctly recalled it was removed from further test periods but presented two more times in the next two study periods. In the Test condition, once a pair was correctly recalled it was removed from further study periods but tested two more times in the next two test periods. In the instructions at the beginning of the experiment the subjects were informed about whether recalled items would be dropped, studied two more times, or tested two more times. Subjects were also informed that they would take a final test (a cued recall test identical to a test period from the learning phase) in the second session 1 week later.

At the end of the learning phase the subjects were asked to predict how many of the 40 word pairs they would correctly recall on the final test in 1 week. The subjects were then dismissed and returned for the final test 1 week later. On the final test the subjects were shown each Swahili word with a cursor below it and were told to type the correct English word for each Swahili word. Each final test trial lasted 15 s (with a 500-ms intertrial interval), and subjects were not required to make JOLs. After completing the final test the subjects were debriefed and thanked for their participation.

## Results

All results unless otherwise stated were significant at the .05 level.

**Cumulative learning.** Because the Drop and Study conditions did not require repeated recall of the entire list during each test period, traditional learning curves were not examined. Instead, Figure 1 shows cumulative recall during the learning phase, which is the proportion of pairs recalled at least once as measured in each test period (see Karpicke & Roediger, 2007, 2008). The figure shows that there were virtually no differences in learning performance across conditions and that more than 95% of the pairs were correctly recalled by the fourth period. Thus, subsequent analyses focus on performance during the first four periods in the learning phase. In the Test condition, recall accuracy changed little across repeated tests, averaging 94.8% and 95.6% on the second and third retrieval attempts, respectively. In other words, intertrial forgetting was minimal in the Test condition (see Tulving, 1964), and therefore cumulative learning performance converged with the traditional method of measuring learning performance.

**Final recall.** The left panel of Figure 2 shows recall on the final test 1 week after learning. Although cumulative learning performance was nearly identical in the three conditions, there was a large effect of repeated retrieval on long-term retention. Restudying recalled pairs two additional times did not produce a signifi-

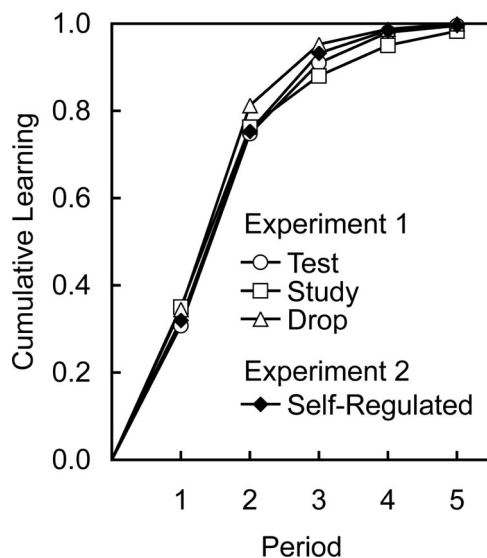


Figure 1. Cumulative learning in the Drop, Study, and Test conditions in Experiment 1 and in the self-regulated condition in Experiment 2.

cant benefit over dropping recalled pairs (.39 vs. .35, respectively,  $F < 1$ ) but repeated retrieval of recalled pairs produced a large gain relative to the other two conditions (.78 vs. .39),  $F(1, 18) = 24.47$ ,  $\eta_p^2 = .58$ ; (.78 vs. .35),  $F(1, 18) = 31.89$ ,  $\eta_p^2 = .64$ . This result is consistent with prior research on the testing effect showing large positive effects of repeated retrieval relative to dropping items (Karpicke & Roediger, 2007, 2008).

Final recall was also examined as a function of different measures of item difficulty. Item difficulty would likely influence subjects' JOLs, and therefore it was important to examine whether item difficulty also influenced the effect of retrieval practice on learning. Table 1 shows final recall as a function of intrinsic difficulty (EOL ratings), encoding fluency, and retrieval fluency. Encoding fluency was determined by performing a median split on each subject's self-paced study time in the period immediately preceding the first correct recall (Koriat & Ma'ayan, 2005). (Due to a program error, self-paced study times were not recorded for one subject, and therefore the analysis of encoding fluency was performed on data from 29 subjects.) Retrieval fluency was determined by a median split on each subject's recall time for the first correct recall (Benjamin et al., 1998). The means of the above- and below-median EOL ratings, study times, and recall times are also shown in Table 1. The data in Table 1 show that each measure of difficulty produced a main effect on final recall: for intrinsic difficulty, .55 vs. .46,  $F(1, 27) = 17.81$ ,  $\eta_p^2 = .40$ ; for encoding fluency, .52 vs. .47,  $F(1, 26) = 3.53$ ,  $\eta_p^2 = .12$ ; for retrieval fluency, .55 vs. .45,  $F(1, 27) = 16.99$ ,  $\eta_p^2 = .39$ . There was also a main effect of learning condition in each analysis: for intrinsic difficulty,  $F(2, 27) = 16.75$ ,  $\eta_p^2 = .55$ ; for encoding fluency,  $F(2, 26) = 15.24$ ,  $\eta_p^2 = .54$ ; for retrieval fluency,  $F(2, 27) = 17.35$ ,  $\eta_p^2 = .56$ . There were no interactions in the analyses of intrinsic difficulty or encoding fluency ( $F_s < 1$ ), but there was an interaction in the analysis of retrieval fluency,  $F(2, 27) = 5.65$ ,  $\eta_p^2 = .30$ , which was due to the fact that there was not a difference between slow and fast retrieval times in the Study condition (.39 vs. .38, respectively). In short, the data in Table 1 show that retrieval practice improved long-term retention of both easy and difficult pairs.

Turning now to the last measure of item difficulty, the number of trials to first correct recall produced a profound effect on final recall (cf. Koriat, 2008). The left panel of Figure 3 shows final recall as a function of the period in which the pair was recalled for the first time (with Test Periods 4–8 combined because the majority of pairs had been recalled by the fourth period). The figure shows that pairs recalled early in learning were more likely to be retained on the final test than pairs recalled later in learning. This pattern occurred in the Drop and Study conditions but was substantially reduced in the Test condition, suggesting again that retrieval practice promoted learning of difficult items. (See Appendix for the results of separate 2 [measure: JOL vs. final recall]  $\times$  4 [period] analyses of variance [ANOVAs] performed on each learning condition.)

**Judgments of learning.** Did subjects' JOLs indicate that they were aware that repeated retrieval practice would enhance retention? The analyses focused on JOLs following the first correct recall of each pair because these JOLs inform the choices subjects make after they can successfully recall an item during learning. In general there were not large differences in the pattern of item-by-item or aggregate JOLs across the three learning conditions, as

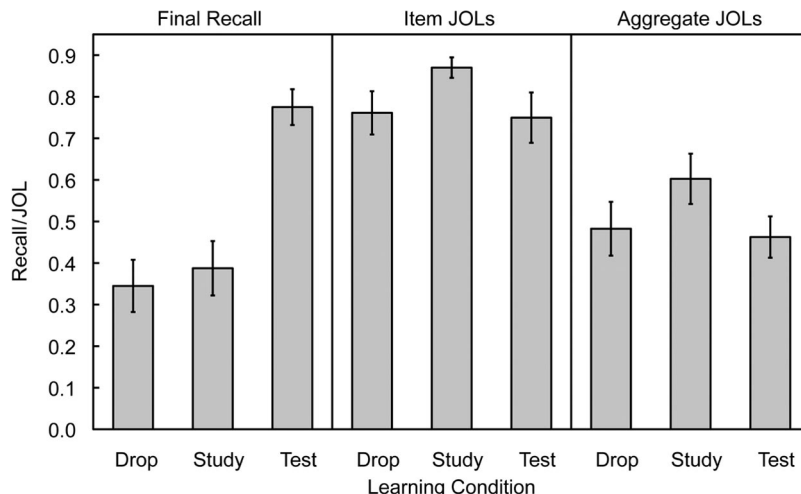


Figure 2. Left panel: Final recall 1 week after the learning phase in the Drop, Study, and Test conditions in Experiment 1. Center panel: Item-by-item judgments of learning (JOLs) following the first correct recall in the learning phase. Right panel: Aggregate JOLs made at the end of the learning phase. Error bars indicate the standard error of the mean.

shown in the center and right panels of Figure 2. Item JOLs averaged .87 in the Study condition, which was greater than those in the Test condition (.76),  $F(1, 18) = 3.56, \eta_p^2 = .17$ , and in the Drop condition (.75),  $F(1, 18) = 3.39, \eta_p^2 = .16$ . The mean JOLs in the Test and Drop conditions were not reliably different ( $F < 1$ ). A similar pattern occurred in the aggregate JOLs made at the end of learning: Subjects in the Study condition predicted they would recall slightly more than half of the pairs on the final test (.60), whereas subjects in the Test and Drop conditions predicted they would recall fewer than half (.48 and .46, respectively). An ANOVA indicated that the aggregate JOLs were not reliably different across the three conditions,  $F(2, 27) = 1.63, ns$ . The slight inflation of JOLs in the Study group is consistent with prior research showing that aggregate predictions of future performance are inflated in repeated study conditions relative to repeated re-

trieval conditions (Roediger & Karpicke, 2006b). Finally, in the Test condition the mean JOLs for correct recalls increased slightly across the first (.76), second (.78), and third (.82) correct recalls,  $F(2, 18) = 4.76, \eta_p^2 = .35$ .

The right panel of Figure 3 shows JOLs as a function of the period in which the pairs were recalled for the first time. These data are in stark contrast to the data shown in the left panel of Figure 3 depicting a decrease in final recall as a function of number of periods to first correct recall. In the Drop condition recall decreased as a function of this factor whereas JOLs increased: interaction,  $F(3, 396) = 3.13, \eta_p^2 = .02$ . Likewise, in the Study condition recall decreased as a function of this factor whereas JOLs showed a slight increase: interaction,  $F(3, 396) = 10.53, \eta_p^2 = .07$ . However, this interaction was not observed in the Test condition ( $F < 1$ ). The overall pattern of results indicates that subjects discounted the number of trials it took them to recall an item as a cue about how well they had learned the item (cf. Koriat, 2008). (Again, see Appendix for the complete analysis of the data presented in Figure 3.)

**Cue utilization: The bases of JOLs.** What factors influenced subjects' JOLs? First, because JOLs were solicited during test periods there was a large influence of recall success on JOLs. Collapsed across conditions and test periods, JOLs for correct recalls averaged .80 whereas JOLs for incorrect trials averaged .03. For our purposes, JOLs for correct recalls are of key importance because of their relevance to the strategy choices that subjects make once they successfully recall information (examined in Experiment 2).

The effects of intrinsic difficulty, number of periods to correct recall, encoding fluency, and retrieval fluency on JOLs following correct recalls were examined by calculating the within-subject gamma correlation between each factor and JOLs. The intrinsic-JOL gamma averaged .23, which was significantly greater than zero,  $F(1, 29) = 22.78, \eta_p^2 = .44$ , whereas the number of periods-JOL gamma averaged .08, which was not significantly different

Table 1  
Final Recall in Experiment 1 as a Function of Intrinsic Difficulty, Encoding Fluency, and Retrieval Fluency

Variable	Mean of difficulty index	Drop	Study	Test
Intrinsic difficulty				
Easy	6.2	.41 (.06)	.42 (.06)	.82 (.05)
Hard	4.1	.28 (.06)	.36 (.08)	.73 (.04)
Encoding fluency (in seconds)				
Fast	4.5	.38 (.07)	.41 (.07)	.81 (.04)
Slow	11.7	.32 (.07)	.37 (.06)	.75 (.05)
Retrieval fluency (in seconds)				
Fast	1.4	.43 (.08)	.39 (.07)	.85 (.04)
Slow	3.1	.26 (.05)	.38 (.07)	.70 (.05)

Note. Standard errors are in parentheses. Intrinsic difficulty = normative ease-of-learning ratings; encoding fluency = self-paced study time on the trial immediately preceding the first correct recall during learning; retrieval fluency = recall time of first correct recall.

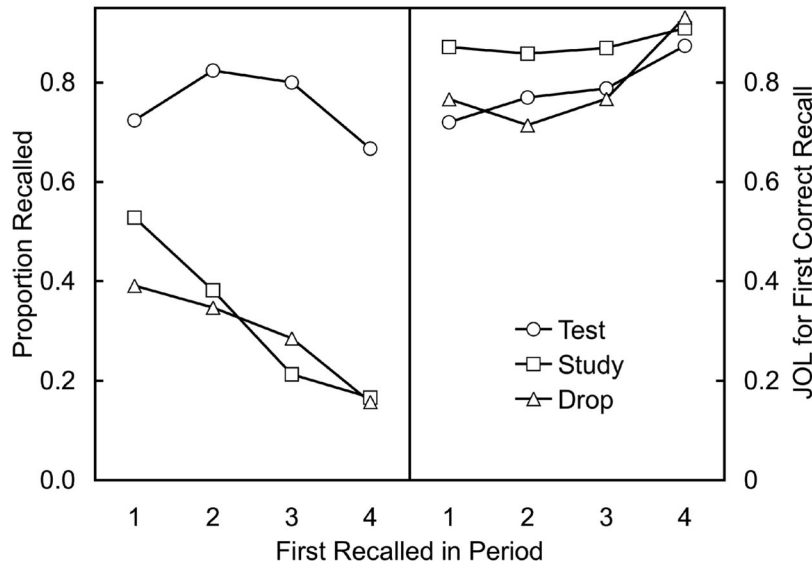


Figure 3. Left panel: Final recall as a function of the period in which the pair was first recalled during learning. Right panel: Mean judgments of learning (JOLs) following the first correct recall of pairs as a function of the period in which they were first recalled.

from zero ( $F < 1$ ). Encoding fluency during study periods produced little effect on JOLs, which was not surprising because JOLs were obtained during test periods. The encoding fluency–JOL gamma averaged  $-.02$ , which was not significantly different from zero ( $F < 1$ ). Retrieval fluency produced the strongest influence on JOLs in test periods. The retrieval fluency–JOL gamma averaged  $-.38$ ,  $F(1, 29) = 58.63$ ,  $\eta_p^2 = .67$ , and the correlation was negative for 28 of the 30 subjects. The negative retrieval fluency–JOL gamma correlation indicates that pairs recalled more quickly were given higher JOLs than were those recalled more slowly (cf. Benjamin et al., 1998).

### Discussion

In Experiment 1 there was a large effect of retrieval practice: Once a word was correctly recalled, repeatedly retrieving it two times produced a large gain in long-term retention while repeated studying produced no benefit relative to dropping it. Yet subjects' JOLs indicated they were not aware that repeated retrieval practice would benefit retention. Retrieval fluency produced the strongest influence on JOLs during test periods—pairs recalled quickly were given higher JOLs than pairs recalled slowly. In sum, the pattern of results suggests that fluent retrieval led to high JOLs the first time items were recalled. Consequently, subjects did not predict that additional retrieval practice would enhance long-term retention, and subjects' JOLs were overconfident relative to long-term retention.

### Experiment 2: Metacognitive Control and Strategy Selection in Test Periods

The implication of Experiment 1 is that despite the powerful effects of retrieval practice on long-term retention, subjects may not choose to practice repeated retrieval because once they can recall an item they believe they have learned it, as evidenced by

their high JOLs. Experiment 2 examined this implication by asking subjects to choose to test, study, or drop items once they had successfully recalled them. The procedure in Experiment 2 was identical to the one in Experiment 1 except that after the first time each pair was recalled the subjects decided whether they wanted to test the pair again, study it again, or remove it from further practice. The subjects were instructed to make their strategy choices in an effort to maximize the number of pairs they would recall on a final test 1 week later. The subjects also made JOLs during test periods, and the relationship between monitoring and strategy choices was examined. The prediction was that fluent retrieval would lead subjects to believe they had learned items, and as a consequence they would often choose to drop items rather than practice retrieval.

### Method

**Subjects, materials, and procedure.** Forty-two Washington University undergraduates participated in Experiment 2 in exchange for payment or course credit. None of the subjects had participated in Experiment 1.

The materials and procedure were identical to those in Experiment 1 in all respects but one: After each pair was recalled for the first time, subjects chose a strategy for the pair. They could choose to continue studying the pair, to repeatedly test the pair, or to remove the pair from further practice. Subjects were informed that if they chose "Study" they would study the pair two more times in the next two study periods but would no longer test the pair. If they chose "Test" they would practice retrieving the pair two more times in the next two test periods but would not study it again. If they chose "Remove" they would not study or test the pair again in any additional periods. After subjects correctly recalled a pair for the first time and after they made their JOL for the pair, the options "1 = Study," "2 = Test," and "3 = Remove" were shown on the computer screen and subjects pressed the 1, 2, or 3 key to

indicate their strategy choice. The subjects were told that there were no right or wrong choices—they could restudy, retest, or remove as many or as few pairs as they wished. They were also told that they did not need to make the same choice for all pairs and could select different strategies for different pairs. However, they were instructed to make their choices with the goal of maximizing the number of pairs they would recall on the final test in 1 week.

At the end of the learning phase the subjects predicted how many of the 40 pairs they would recall on the final test. They were also asked to explain how they made their decisions to study, test, or remove pairs—the subjects wrote down their responses to this question on a sheet of paper provided to them. Subjects took the final test 1 week later.

## Results

**Cumulative learning.** Cumulative learning performance for all 42 subjects is plotted in Figure 1 with the data from Experiment 1. Cumulative learning in Experiment 2 was virtually identical to learning in Experiment 1. The figure shows that the majority of the pairs (more than 98%) were recalled by the fourth period.

**Final recall, judgments of learning, and strategy choices.** Final recall for all 42 subjects averaged .46, which was below the level achieved in the Test condition in Experiment 1 (.78),  $F(1, 50) = 15.62$ ,  $\eta_p^2 = .24$ . Of course, this is a cross-experiment comparison, and it should be interpreted with caution. The left panel of Figure 4 shows final recall for pairs that subjects assigned to the Drop, Study, and Test conditions. Because subjects assigned pairs to different conditions and because not all subjects used all three strategies, the following analyses were performed with pairs as the random variable. Figure 4 shows that practicing retrieval improved retention relative to additional studying (.72 vs. .47),  $F(1, 39) = 54.03$ ,  $\eta_p^2 = .58$ , and dropping items (.72 vs. .35),  $F(1, 39) = 294.63$ ,  $\eta_p^2 = .88$ . There was also a benefit of repeated study versus dropping items (.47 vs. .35),  $F(1, 39) = 14.79$ ,  $\eta_p^2 = .28$ .

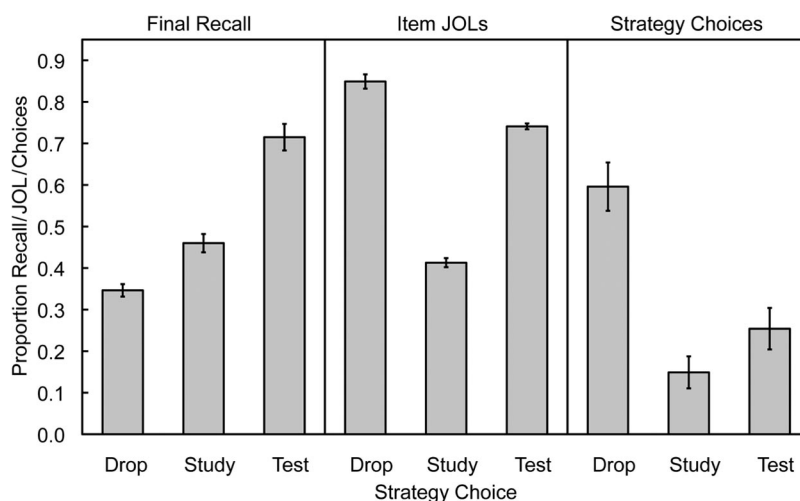
The center panel of Figure 4 shows mean item-by-item JOLs made after the first correct recall of each pair. JOLs averaged .85 for pairs that were subsequently dropped, .74 for pairs that were tested, and .41 for pairs that were studied,  $F(2, 78) = 527.12$ ,  $\eta_p^2 = .93$ . The result differs from Experiment 1 probably because JOLs were a consequence of the strategy given to subjects in Experiment 1, whereas JOLs determined the strategy chosen by subjects in Experiment 2 (but again, caution should be exercised in interpreting this cross-experiment comparison). For pairs that subjects chose to test the mean JOLs increased across the first (.74), second (.82), and third (.87) correct recalls,  $F(2, 78) = 68.72$ ,  $\eta_p^2 = .64$ .

Aggregate JOLs made at the end of the learning phase averaged .464, which was almost identical to the average level of final recall (.457). However, there was only a modest correlation between aggregate JOLs and final recall across subjects ( $r = .31$ ). Thus a subject's aggregate JOL was not necessarily an accurate predictor of his or her final recall.

The right panel of Figure 4 shows the proportion of pairs assigned to each strategy condition by all 42 subjects. The majority of the pairs were dropped (.60), fewer pairs were repeatedly tested (.25), and even fewer pairs were repeatedly studied (.15). The metacognitive control processes involved in strategy selection are discussed in a later section.

**Cue utilization: The bases of JOLs.** As in Experiment 1, there was a large influence of recall success on JOLs. Collapsed across conditions and test periods, JOLs for correct recalls averaged .76, whereas JOLs for incorrect trials averaged .06. The present analysis focuses on the bases of JOLs for correct recalls because these JOLs immediately preceded subjects' strategy choices.

The average within-subject gamma correlations between four factors (intrinsic difficulty, number of periods to correct recall, encoding fluency, and retrieval fluency) and JOLs following correct recalls were calculated for 41 subjects (one subject gave the same JOL rating to all pairs and was therefore excluded from the



*Figure 4.* Left panel: Final recall 1 week after the learning phase in the Drop, Study, and Test conditions in Experiment 2. Center panel: Item-by-item judgments of learning (JOLs) following the first correct recall in the learning phase. Right panel: Proportion of times that subjects chose to drop, study, or test items. Error bars indicate the standard error of the mean.

analysis). The pattern of correlations replicated the results of Experiment 1. The intrinsic-JOL gamma averaged .26,  $F(1, 40) = 8.99$ ,  $\eta_p^2 = .18$ , whereas the number of periods-JOL gamma averaged .06,  $F(1, 40) = 3.70$ ,  $\eta_p^2 = .08$ . Encoding fluency produced little influence on JOLs, with the encoding fluency-JOL gamma averaging  $-.04$  ( $F < 1$ ). Retrieval fluency produced the strongest influence on JOLs. The retrieval fluency-JOL correlation averaged  $-.37$ ,  $F(1, 40) = 100.60$ ,  $\eta_p^2 = .72$ , and the correlation was negative for 38 of the 41 subjects. Subjects generally relied on retrieval fluency when making JOLs during test periods—they gave higher JOLs to pairs recalled quickly than to ones recalled slowly.

*Effects of strategy choices on final recall.* As noted earlier, at the group level subjects tested themselves on 25% of the pairs, restudied 15% of the pairs, and dropped the majority of the pairs (60%). However, subjects differed in which strategies they were most likely to select, as shown in Figure 5. The subjects were divided based on the percentage of their drop, study, and test choices (using six bins: 0, 1–20, 21–40, 41–60, 61–80, and 81–100). The figure shows final recall as a function of the percentage of the strategy choices and also shows the number of subjects within each bin. There was a positive relationship between the percentage of test choices and final recall ( $r = .66$ ), essentially zero relationship between the percentage of restudy choices and final recall ( $r = .01$ ), and a negative relationship between the percentage of drop choices and final recall ( $r = -.58$ ). Thus the strategies subjects selected had important consequences for final recall.

*Strategy modulation: Theory-based versus experience-based strategy selection.* It is clear from the data in Figure 5 that subjects differed in the degree to which they modulated their strategies. Some subjects selected a preferred strategy for nearly all pairs, probably on the basis of a theory of what strategy would promote learning. Other subjects applied different strategies to different pairs on the basis of their experience in the task. The subjects were divided into two groups based on how frequently they chose a single strategy. Eighteen subjects selected just one strategy for at least 90% of the pairs. These subjects were labeled *theory-based subjects* (cf. Koriat, Bjork, Sheffer, & Bar, 2004) based on the the idea that these subjects relied on an a priori theory

that they applied to nearly all pairs (e.g., “If I can recall it I should drop it because I know it” or “I should test myself on all pairs because that will help me remember them in a week”). Of the 18 theory-based subjects, 3 subjects chose to test more than 90% of the pairs, 1 subject chose to restudy more than 90% of the pairs, and the remaining 14 subjects chose to drop more than 90% of the pairs. Ten subjects dropped 100% of the pairs.

The other 24 subjects were labeled *experience-based subjects* based on the idea that these subjects relied on their experience during learning (their JOLs) to modulate their strategy choices. Two analyses provide converging evidence for the distinction between theory-based and experience-based strategy selection. First, strategy choice response times averaged 1,125 ms for theory-based subjects and 1,816 ms for experience-based subjects,  $F(1, 40) = 13.51$ ,  $\eta_p^2 = .25$ , a 61% increase in choice response times in the experience-based group. Second, further evidence comes from subjects’ self-reports solicited at the end of the learning phase in Session 1. Sixteen of the 18 theory-based subjects reported that they applied a single strategy to the majority of the pairs and that their strategy selection was not related to their JOLs. None of the 18 subjects indicated that they based their strategies on their assessments of their current learning. In contrast, 22 of 24 experience-based subjects stated that they selected strategies based on their assessment of their confidence in their own learning or their feeling of how well they knew the pair or the amount of time it took them to recall the target word. Thus there was converging evidence that some subjects selected strategies based on a theory of what they believed would promote learning while others modulated their strategy choices based on their experience in the learning task.

The top portion of Table 2 shows the mean proportion of choices for the experience-based and theory-based subjects. By definition the theory-based subjects selected a single strategy for more than 90% of the pairs, and the data show that this was the case for subjects who decided to drop, study, or test on the majority of the pairs. Table 2 also indicates that the experience-based subjects were more likely to drop pairs than study them,  $F(1, 23) = 6.85$ ,  $\eta_p^2 = .23$ , or test them (this difference approached significance),  $F(1, 23) = 2.13$ ,  $\eta_p^2 = .09$ ,  $p = .08$ . The proportion of study and test choices were not significantly different,  $F(1, 23) = 1.31$ , *ns*.

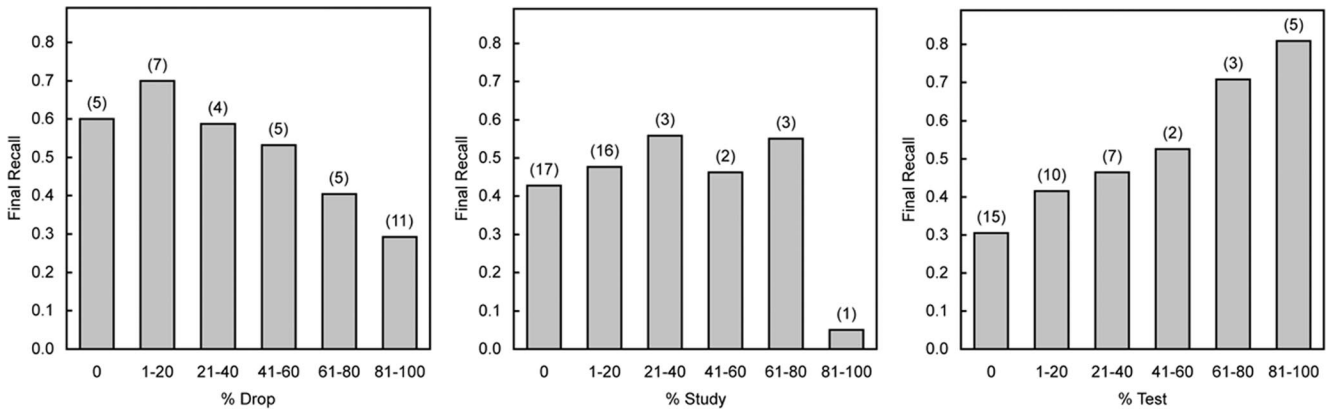


Figure 5. Final recall as a function of the percentage of Drop, Study, and Test choices. Numbers of subjects are in parentheses.



Table 2  
Mean Proportion of Strategy Choices and Mean JOLs for Each Strategy Choice in Experiment 2

Variable	Strategy choice		
	Drop	Study	Test
Proportion of choices			
Experience-based	.47 (.06)	.22 (.05)	.32 (.06)
Theory: drop	.99 (.01)		
Theory: study		.93 (—)	
Theory: test			.98 (.01)
JOLs			
Experience-based	.83 (.04)	.48 (.05)	.68 (.03)
Theory: drop	.82 (.05)		
Theory: study		.24 (—)	
Theory: test			.85 (.04)

Note. JOLs = judgments of learning; Experience-based  $n = 24$ ; theory: drop  $n = 14$ ; theory: study  $n = 1$ ; theory: test  $n = 3$ . Standard errors are in parentheses. Dashes indicate that no standard error is reported because  $n = 1$ .

The bottom portion of Table 2 shows the mean JOLs associated with each strategy choice. The theory-based subjects who dropped or tested the majority of pairs exhibited high levels of confidence in their learning, whereas the one subject who studied the majority of pairs reported rather low JOLs. For the remaining experience-based subjects JOLs were highest for pairs that were subsequently dropped. The JOLs for drop choices were greater than JOLs for test choices,  $F(1, 19) = 30.55$ ,  $\eta_p^2 = .62$ , and study choices,  $F(1, 17) = 29.20$ ,  $\eta_p^2 = .63$ . JOLs for test choices were greater than JOLs for study choices,  $F(1, 17) = 31.98$ ,  $\eta_p^2 = .65$ .

*A model of metacognitive control and strategy selection.* The last analysis was carried out to see whether a schematic model of subjects' decision rules would fit the strategy choice data. The model assumes that subjects set two decision criteria: a criterion for dropping or practicing a pair (labeled  $D_C$ ) and, for practiced pairs, a criterion for testing or studying the pair (labeled  $T_C$ ). Subjects first decide whether to drop the current pair or continue practicing it. This assumption is supported by decision latency data showing that among the experience-based subjects the decision to drop an item was made faster than the decision about how to practice it (1,550 ms vs. 1,911 ms),  $F(1, 23) = 3.36$ ,  $\eta_p^2 = .13$ . The model also assumes that if subjects' assessments of their current learning (their JOLs) meet or exceed  $D_C$  then they will drop pairs, and otherwise they will continue to practice them. For pairs that subjects practice, if subjects' JOLs meet or exceed  $T_C$  then they will test themselves, and otherwise they will study pairs. The model accounts for differences between theory-based and experience-based strategy selection in terms of where subjects set their criteria. For example, a theory-based subject who dropped most of the pairs would set a low  $D_C$  so that most JOLs would fall above the criterion and would be dropped. Likewise, subjects who tested themselves on most pairs would set a relatively high  $D_C$  and a relatively low  $T_C$ , and subjects who studied most pairs would set both  $D_C$  and  $T_C$  relatively high.

The relationship between subjective experience (JOLs) and strategy selection was assessed as the within-subject gamma correlation (a) between JOLs and the decision to drop or practice pairs and (b) for practiced pairs between JOLs and the decision whether

to test or to study the pairs. The JOL-Drop gamma averaged .87 (based on data from 26 subjects), and the JOL-Test gamma averaged .81 (based on data from 18 subjects). The correlations indicate a strong link between subjects' assessments of their own learning and their strategy selection.

A computational procedure was used to calculate estimates of  $D_C$  and  $T_C$  for each subject. The strategy selection criteria were calculated by considering each JOL rating actually used by each subject as a candidate criterion. In calculating  $D_C$ , the proportion of pairs dropped above the candidate and practiced below the candidate was considered, and the overall proportion of pairs above and below the candidate criterion was deemed the fit ratio. The same procedure was carried out for practiced pairs to estimate  $T_C$ , with the overall proportion of pairs tested above and studied below the candidate criterion assessed as the fit ratio. For each subject the candidate criteria that maximized each fit ratio were considered the estimates of  $D_C$  and  $T_C$ . The criterion estimation technique was derived from Koriat and Goldsmith's (1996) method of calculating the response criterion in their analysis of the metacognitive control processes involved in memory reporting.

For the 24 experience-based subjects the mean  $D_C$  estimate was .88 (fit ratio = .83) and was greater than the mean  $T_C$  estimate, which was .56 (fit ratio = .86),  $F(1, 23) = 15.37$ ,  $\eta_p^2 = .40$ . It is also worth noting that for the 14 theory-based subjects who dropped the majority of the pairs the mean  $D_C$  was .25 (fit ratio = .99), which is consistent with the idea that these subjects set a low criterion for dropping the pairs. Likewise, for the three subjects who tested most pairs the mean  $D_C$  was .98 (fit ratio = .98) and the mean  $T_C$  was .10 (fit ratio = .99), consistent with the idea that these subjects set a high criterion for dropping pairs and a low criterion for testing them.

## Discussion

In Experiment 2 the subjects were given control over their own learning: Once they successfully recalled an item they could choose to restudy it, repeatedly test it, or remove it from practice. The analyses of memory performance and metacognitive monitoring replicated essentially all the results observed in Experiment 1 wherein learning conditions were experimentally controlled. Repeated retrieval practice produced superior retention relative to repeated studying or dropping of pairs. Subjects' JOLs exhibited a general tendency toward overconfidence, especially for pairs they chose to drop, and retrieval fluency produced the strongest influence on JOLs solicited in test periods. Thus a cascade of effects was responsible for performance in the learner-controlled conditions in Experiment 2: (a) Fluent retrieval during the learning phase led subjects to give high JOL ratings, (b) JOLs influenced the strategies that subjects selected, (c) higher JOLs led subjects to drop pairs from further practice rather than practice retrieval (or restudy), and (d) dropping pairs after they were recalled only once produced poor long-term retention.

### Experiment 3: The Potentiating Effect of Retrieval on Learning

In Experiments 3 and 4 the focus shifts to strategy choices during study periods. The purpose of Experiment 3 was to estab-

lish the potentiating effect of retrieval on learning and to collect benchmark data as a reference point for Experiment 4 wherein learning conditions were subject-controlled. The potentiating effect of retrieval refers to the finding that subjects acquire more new pairs in a study period following repeated tests versus a single test, even when subjects receive no feedback following the tests and performance changes little across tests (see Izawa, 1969; Karpicke & Roediger, 2007). Thus the act of attempting retrieval improves subjects' encoding of new pairs in the subsequent study period. The implication of the potentiating effect is that inserting retrieval attempts early in the learning phase—even though recall would be less than perfect—would facilitate performance more than additional studying without testing.

Experiment 3 examined three learning conditions. In the standard study/test condition, subjects studied and recalled the entire list of vocabulary word pairs in alternating study and test periods for a total of six periods (STSTST). A second group of subjects studied the list in three study periods before the first test, then studied again and tested again (SSSTST). A third group studied the list in five study periods before the first test (SSSSST), testing only once in the sixth and final period of the learning phase. Note that in contrast to the previous experiments, subjects studied or tested on the entire list in every period—no pairs were removed from study or test periods. The prediction was that attempting retrieval during the learning phase would promote better learning than repeated studying. Thus the STSTST condition should produce the best learning even though subjects in the three conditions were exposed to the pairs for the same total number of trials and periods.

### Method

*Subjects, materials, and procedure.* Forty-eight Washington University undergraduates participated in Experiment 3 in exchange for payment or course credit. None of the subjects had participated in the prior experiments.

Sixty Swahili–English word pairs were used in Experiment 3. The EOL ratings for the 60 pairs averaged 4.9 (median = 4.8,  $SD = 1.2$ ) and ranged from 2.6 to 7.8. On the basis of a median-split on the EOL ratings, 30 were deemed easy pairs ( $M = 6.0$ ) and 30 were deemed difficult ( $M = 4.0$ ).

The procedure was similar to the ones used in the previous experiments except that subjects made JOLs in study periods, not test periods. In the instructions at the beginning of the experiment the subjects were informed about the nature of their particular learning condition. In the STSTST condition, the subjects were told that they would study and test on the list in alternating study and test periods for a total of six periods. In the SSSTST condition, the subjects were told that they would study the list in three study periods before the first test and then would study the list again before the second test. In the SSSSST condition, the subjects were told that they would study the list in five study periods before the first and only test. The order of trials within each period was randomized by the computer.

Finally, the subjects in all conditions were instructed not to test themselves during the study periods (e.g., by closing their eyes or looking away from the screen or covering up the English word and trying to recall it). The experimenter monitored compliance with this instruction. This instruction was included because during preliminary pilot testing some subjects in the repeated study con-

ditions began to use such strategies to test themselves during study periods. Of course, subjects' decisions to self-test were examined in Experiment 4.

### Results

*Learning performance: The potentiating effect of retrieval.* The key result of Experiment 3 is shown in Figure 6, which shows the mean proportion of pairs recalled in each test period (traditional learning curves). The dashed line connects the first test in the three conditions and shows that performance increased across repeated study periods. Recall increased from .25 to .52 to .61 after Study Periods 1, 3, and 5, respectively,  $F(2, 45) = 11.50$ ,  $\eta_p^2 = .34$ . More importantly, Figure 6 shows the potentiating effect of testing on learning. On the test in Period 4 the STSTST condition outperformed the SSSTST condition by about 11% (.63 vs. .52). This difference was marginally significant,  $F(1, 30) = 1.77$ ,  $\eta_p^2 = .06$ ,  $p = .09$ . Performance on the test in Period 6 increased as a function of the number of prior tests from .61 to .84 in the SSSSST, SSSTST, and STSTST conditions, respectively,  $F(2, 45) = 5.00$ ,  $\eta_p^2 = .18$ . Thus inserting retrieval attempts improved performance during the learning phase.

Because subjects may use the perceived intrinsic difficulty of items as a cue for their JOLs and strategy choices, it is important to examine whether the potentiating effect of retrieval depends on item difficulty. Table 3 shows learning performance for easy and hard pairs as determined by EOL ratings. It is possible that retrieval might potentiate the learning of easy pairs but that additional studying might promote learning of hard pairs. This was not the case—the potentiating effect of retrieval occurred for both easy and hard pairs. In Period 4 the STSTST group recalled more than the SSSTST group (marginally significant),  $F(1, 30) = 1.79$ ,  $\eta_p^2 = .06$ ,  $p = .09$ ; both groups recalled more easy pairs than hard pairs,  $F(1, 30) = 67.86$ ,  $\eta_p^2 = .69$ ; and there was no Condition  $\times$

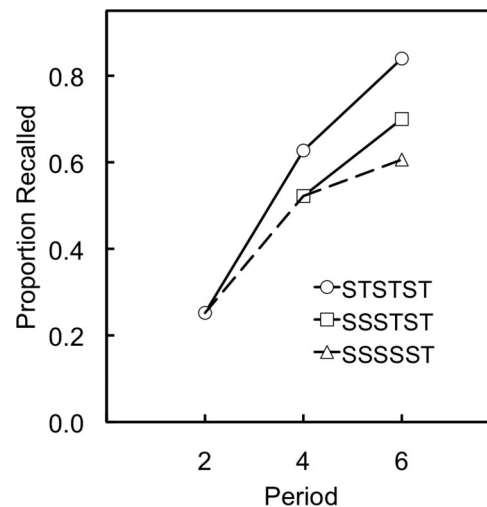


Figure 6. The potentiating effect of testing on learning. Mean proportion of word pairs recalled in each test period in the STSTST, SSSTST, and SSSSST conditions in Experiment 3. The dashed line connects performance on the first test across conditions to show the effect of repeated studying on recall. The solid lines connect performance within each condition. S = study period; T = test period.

Table 3  
*Learning Performance in Experiment 3 for Intrinsically Easy and Hard Word Pairs, on the Basis of a Median Split on Ease-of-Learning Ratings*

Word pair difficulty and learning condition	Period 2	Period 4	Period 6
Easy pairs			
STSTST	.30 (.05)	.72 (.05)	.91 (.03)
SSSTST		.62 (.06)	.77 (.05)
SSSSST			.69 (.06)
Hard pairs			
STSTST	.21 (.05)	.55 (.06)	.78 (.05)
SSSTST		.44 (.06)	.64 (.06)
SSSSST			.53 (.07)

Note. Standard errors are in parentheses. S = study period; T = test period.

Easy/Hard interaction ( $F < 1$ ). In Period 6 there was a main effect of condition,  $F(1, 45) = 5.03$ ,  $\eta_p^2 = .18$ , and a main effect of easy versus hard pairs,  $F(1, 45) = 69.84$ ,  $\eta_p^2 = .61$ , and no interaction ( $F < 1$ ). Thus retrieval attempts potentiated learning of both inherently easy and inherently hard pairs.

*Judgments of learning.* Figure 7 shows the mean JOLs in each study period compared to the mean proportion recalled in each test period. The top panel shows the STSTST condition and a pattern of results called the underconfidence-with-practice effect (Koriat, Sheffer, & Ma'ayan, 2002). JOLs in Study Period 1 were greater than the mean proportion recalled in Test Period 2 (.43 vs. .28),  $F(1, 15) = 7.02$ ,  $\eta_p^2 = .32$ , indicating overconfidence in Period 1. However, JOLs in Study Period 3 underestimated recall in Period 4 (.48 vs. .68),  $F(1, 15) = 12.20$ ,  $\eta_p^2 = .45$ , and the underconfidence effect occurred in Periods 5/6 as well (.72 vs. .88),  $F(1, 15) = 23.72$ ,  $\eta_p^2 = .61$ .

The underconfidence-with-practice effect also occurred following repeated study periods in the other two conditions. In the SSSTST condition, JOLs in Period 3 were lower than recall in Period 4 (.45 vs. .52; this difference approached significance),  $F(1, 15) = 1.68$ ,  $\eta_p^2 = .10$ ,  $p = .10$ . The pattern of underconfidence occurred in Periods 5/6 in the SSSTST condition (.58 vs. .70),  $F(1, 15) = 8.72$ ,  $\eta_p^2 = .37$ , and in the SSSSST condition (.51 vs. .61),  $F(1, 15) = 2.57$ ,  $\eta_p^2 = .15$ .

*Changes in the bases of JOLs across periods.* The last analysis examined the cues that subjects relied on as bases for their JOLs and how those cues may have changed across periods. Specifically, Koriat (1997) proposed that subjects rely on cues like the perceived intrinsic difficulty of items early in learning (in the first study period) but rely on internal mnemonic cues like encoding fluency later in learning. Figure 8 depicts this shift in the cues used for JOLs. The left ordinate shows the mean gamma correlations between intrinsic difficulty (EOL ratings) and JOLs, with positive correlations representing greater reliance on intrinsic cues. The right ordinate shows the mean gamma correlations between encoding fluency (measured as self-paced study times) and JOLs, with negative correlations indicating greater reliance on encoding fluency (cf. Koriat et al., 2006). The top panel shows the STSTST condition. The intrinsic-JOL gammas decreased across periods,  $F(2, 30) = 9.64$ ,  $\eta_p^2 = .39$ , whereas the encoding fluency-JOL gammas increased across periods,  $F(2, 30) = 81.88$ ,  $\eta_p^2 = .85$ . In

Period 1 the intrinsic-JOL gamma averaged .25 while the encoding fluency-JOL gamma averaged .11, indicating greater reliance on intrinsic difficulty than on encoding fluency,  $F(1, 15) = 4.31$ ,  $\eta_p^2 = .22$ . However, the pattern changed in Period 3 with subjects showing less reliance on intrinsic difficulty and greater reliance on encoding fluency,  $F(1, 15) = 22.45$ ,  $\eta_p^2 = .60$ . The intrinsic-JOL correlation remained near zero in Period 5 while the encoding fluency-JOL correlation increased,  $F(1, 15) = 84.89$ ,  $\eta_p^2 = .85$ .

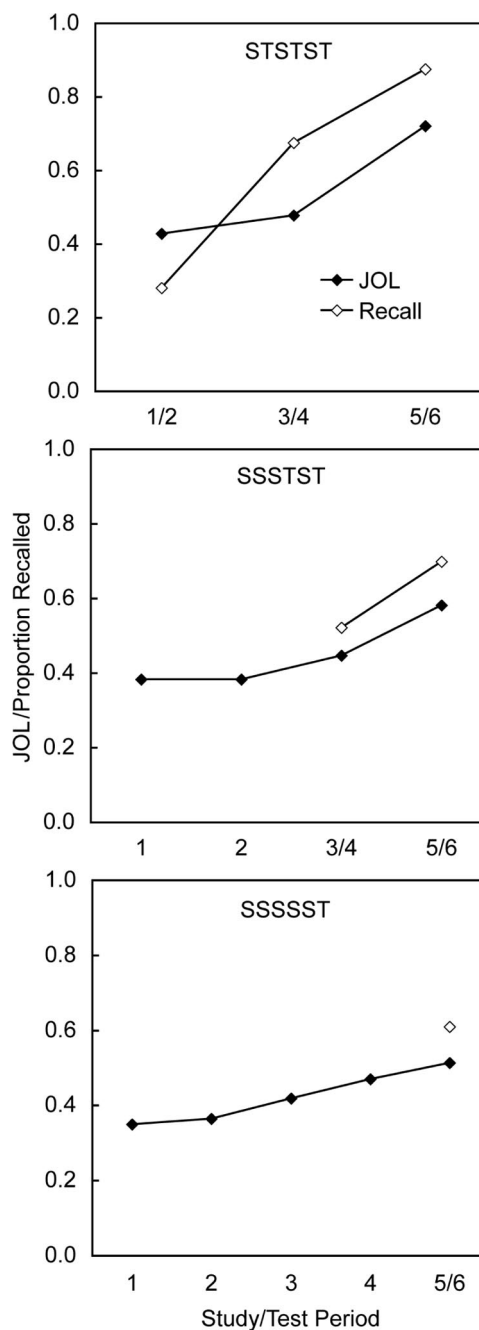


Figure 7. Mean judgments of learning in each study period (S) and mean proportion recalled in each test period (T) in Experiment 3, demonstrating the underconfidence-with-practice effect. JOL = judgment of learning.

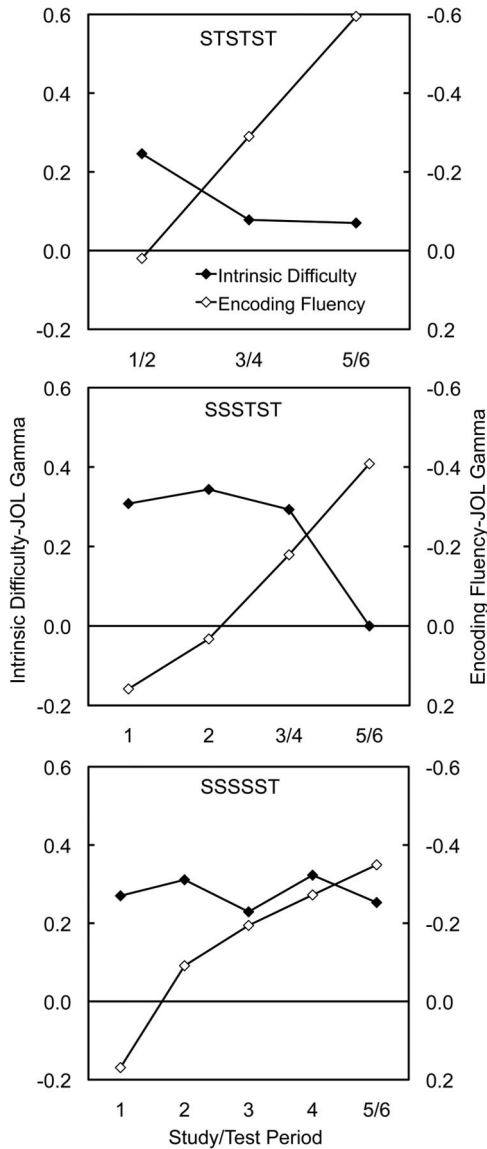


Figure 8. Gamma correlations between intrinsic difficulty (ease-of-learning ratings) and judgments of learning (JOLs; plotted on the left ordinate) and encoding fluency and JOLs (plotted on the right ordinate) in Experiment 3, showing a shift in the bases of JOLs across periods. S = study period; T = test period.

The middle and bottom panels of Figure 8 show the gamma correlations in the SSSTST and SSSSST conditions. In the SSSTST condition the intrinsic-JOL gammas decreased across periods,  $F(3, 45) = 10.35$ ,  $\eta_p^2 = .41$ , while the encoding fluency-JOL gammas increased across periods,  $F(3, 45) = 23.54$ ,  $\eta_p^2 = .61$ . In the SSSSST condition there was no change in the intrinsic-JOL gammas across periods,  $F(4, 60) = 1.01$ , *ns*, while again the encoding fluency-JOL gammas increased across periods,  $F(4, 60) = 16.31$ ,  $\eta_p^2 = .52$ . Both conditions showed slightly greater reliance on intrinsic difficulty relative to encoding fluency in Period 1: SSSTST:  $F(1, 13) = 4.72$ ,  $\eta_p^2 = .27$ ; SSSSST:  $F(1, 15) = 2.37$ ,  $\eta_p^2 = .14$ . However, across repeated study periods,

subjects continued to rely on intrinsic difficulty as the basis for JOLs. In the SSSTST condition the shift toward greater reliance on encoding fluency did not occur until the fifth study period following a Period 4 test,  $F(1, 15) = 40.78$ ,  $\eta_p^2 = .73$ . The shift never occurred in the SSSSST condition—subjects in this condition did not differ in their reliance on intrinsic difficulty as a basis for JOLs across the five study periods ( $F < 1$ ). In sum, the data in Figure 8 show that the shift from reliance on intrinsic difficulty to reliance on internal mnemonic cues depended on testing.

Discussion

The key result of Experiment 3 was that inserting retrieval attempts in the learning phase produced more learning than spending additional time studying, as shown in Figure 6. Thus attempting retrieval potentiated learning in subsequent study periods. In addition, testing produced a shift in the information that subjects used as a basis for JOLs. When subjects repeatedly studied without testing they continued to rely on intrinsic difficulty as a cue for JOLs, but when subjects studied and tested in alternating periods they shifted toward greater reliance on internal mnemonic cues (i.e., encoding fluency) as the learning phase progressed (cf. Koriat, 1997).

Experiment 4: Metacognitive Control and Strategy Selection in Study Periods

The implication of Experiment 3 is that attempting retrieval potentiates learning by enhancing encoding during subsequent study. Experiment 4 examined whether students would choose to attempt retrieval even if they were not sure they could recall an item or if instead they would wait until they reached a sufficient level of confidence in their learning before attempting retrieval. The procedure in Experiment 4 was similar to the one in Experiment 3 except that after studying each word pair the subjects indicated whether they wanted to study the pair again in the next study period, test the pair at the end of the study period (and study it again in the next study period), or remove the pair from further practice. This experiment makes it possible to examine the relationship between metacognitive monitoring and strategy choices made during study periods. The prediction was that rather than attempting retrieval of all items early in learning, subjects would wait until their JOLs reached a sufficient level before they chose to test themselves—despite the fact that the act of attempting retrieval itself would potentiate and improve learning.

Method

Subjects, materials, and procedure. Thirty Washington University undergraduates participated in Experiment 4 in exchange for payment or course credit. None of the subjects had participated in the prior experiments.

The list of 60 Swahili-English pairs used in Experiment 3 was used again in Experiment 4, and the procedure was similar to the one in Experiment 3 the subjects learned the list of Swahili-English pairs in a series of study and test periods, and at the beginning of the experiment the subjects were told that they would have the opportunity to study the list in a possible total of four

study periods. Study trials were self-paced, and JOLs were solicited after each study trial.

Subjects made their strategy choices during study periods. After studying and making their JOL for each pair the subjects were asked to select a strategy. The three strategy options were the same as those in Experiment 2: Subjects could continue studying the pair, test the pair (and also study it again), or remove the pair from further practice. Subjects were informed that if they chose “Study,” the pair would remain in the list during the next study period (and therefore they would have another chance to select another strategy for the pair in the next period). If they chose “Test,” the pair would appear during a test period at the end of the current study period (and in addition the pair would remain in the list in the next study period so subjects could study it again and make a strategy choice again). If subjects chose to remove a pair, it was dropped from further study periods (and therefore from test periods). The procedure for soliciting subjects’ strategy choices was the same as in Experiment 2. Finally, the subjects were told that in the last (fourth) study period they should make their strategy choices as if there were going to be another study phase. That is, if they wished to study the pair again they should choose “Study” even though there would not be another study period. If they chose “Test” they were tested on the pair at the end of the fourth study period.

**Results**

*Learning performance.* Table 4 shows performance during the learning phase. The first row shows the cumulative proportion of the list recalled across the four periods in the learning phase. The cumulative learning data indicate that subjects on average did not recall all pairs in the learning phase. Indeed, cumulative learning appeared to be worse after three study periods (.65) than it was after the same number of study periods in the STSTST condition of Experiment 3 (.85),  $F(1, 44) = 7.03, \eta_p^2 = .14$ . Even with an additional fourth study period in the present experiment the subjects did not recall as many items as did the subjects in Experiment 3 following three study periods (.73 vs. .85, respectively),  $F(1, 44) = 2.58, \eta_p^2 = .06$ . Once again, caution should be exercised in interpreting these cross-experiment comparisons. The bottom portion of Table 4 shows the proportion of words tested in each period (the proportion attempted out of 60), the proportion recalled (out of 60), and the proportion of those tested that were recalled (recalled/

attempted). These data are presented simply to illustrate that cumulative and traditional measures of multitrial learning produced similar patterns of results in Experiment 4.

It is also worth examining whether attempting recall promoted learning even when the recall attempts failed—as was the case in Experiment 3. Table 5 shows the probability of recalling a word pair in Period 2, 3, or 4 as a function of whether the pair was tested but not recalled in the previous period (a failed recall attempt) or not tested (meaning subjects chose to study it again rather than attempt recall). The analysis was performed with pairs as the random variable. The analyses in Periods 2 and 3 were based on 60 pairs and the analysis in Period 4 was based on 48 pairs, because due to the increase in dropping across periods there were 12 pairs without data by Period 4. Of course there are item selection effects present in this analysis because subjects chose which pairs to test, but even so the data replicate the key result of Experiment 3 and show that failed recall attempts potentiated learning relative to not attempting recall, Period 2:  $F(1, 59) = 4.88, \eta_p^2 = .08$ ; Period 3:  $F(1, 59) = 29.99, \eta_p^2 = .34$ ; Period 4:  $F(1, 47) = 6.66, \eta_p^2 = .12$ . Thus attempting recall—even when the recall attempt failed—improved learning in a subsequent study period relative to studying the pairs without testing them (cf. Kornell et al., 2009).

*Judgments of learning.* Item-by-item JOLs averaged .83 for pairs that were dropped, .58 for pairs that were tested, and .40 for pairs that were studied,  $F(2, 118) = 2,958.35, \eta_p^2 = .98$ . This pattern of JOLs is the same pattern observed in Experiment 2 and is consistent with the schematic model of strategy selection described earlier.

*Changes in the bases of JOLs across periods.* The next analysis examined changes in the bases of JOLs across periods as a function of whether pairs were tested in the previous period. Figure 9 shows the average gamma correlations between intrinsic difficulty and JOLs plotted on the left ordinate and between encoding fluency and JOLs plotted on the right ordinate. The top panel of Figure 9 shows the correlations for pairs that were tested and depicts a shift in the basis of JOLs across periods from greater reliance on intrinsic difficulty in Period 1 to increased reliance on encoding fluency in Periods 2–4 (cf. Figure 9 to Figure 8). There was no change in the intrinsic-JOL gammas decreased across periods,  $F(3, 45) = 1.01, ns$ , while the encoding fluency–JOL gammas increased across periods,  $F(3, 45) = 22.80, \eta_p^2 = .60$ . However, the bottom panel of Figure 9 shows that the shift in the basis of JOLs did not occur for pairs that were studied but not tested. There were no changes in the intrinsic-JOL or encoding fluency–JOL gammas across periods ( $F_s < 1$ ). Thus when subjects repeatedly studied the pairs they continued to rely on intrinsic difficulty and showed little reliance on the internal mnemonic cue of encoding fluency. The pattern of results conceptually replicates the results of Experiment 3 wherein repeated study versus alternating study/test periods was experimentally manipulated.

*Strategy choices across study periods.* In contrast to the procedure in Experiment 2, in this experiment subjects could select multiple strategies for the same pairs across periods. For example, if a subject chose “Study” or “Test” for a pair in Period 1 then the pair would appear again in Period 2 and subjects could select another strategy for the same pair. Figure 10 shows the proportion of pairs that were studied, tested, or dropped in each period, with the proportions calculated as the number times each strategy was chosen divided by the total number of pairs that had not yet been

Table 4  
Recall Performance During the Learning Phase in Experiment 4

Variable	Period			
	1	2	3	4
Cumulative				
Recalled	.19 (.04)	.46 (.05)	.65 (.05)	.73 (.05)
Attempted	.38 (.06)	.68 (.05)	.80 (.04)	.87 (.03)
Recalled/attempted	.47 (.05)	.67 (.04)	.80 (.03)	.82 (.03)
Total				
Recalled	.19 (.04)	.35 (.03)	.34 (.05)	.24 (.05)
Attempted	.38 (.06)	.50 (.04)	.41 (.06)	.30 (.05)
Recalled/attempted	.47 (.05)	.70 (.03)	.83 (.03)	.81 (.05)

Note. The data are proportions, with standard errors in parentheses.

Table 5  
Proportion Correctly Recalled in Period *n* Following a Failed Recall Attempt or Additional Study in Period *n*-1

Variable	Period <i>n</i>		
	2	3	4
Correct in period <i>n</i> following failed recall in period <i>n</i> -1	.41 (.03)	.60 (.03)	.50 (.06)
Correct in period <i>n</i> following no test (i.e., studying) in period <i>n</i> -1	.35 (.02)	.42 (.02)	.34 (.01)

Note. Standard errors are in parentheses.

dropped. The figure shows that subjects exhibited a shift in the pattern of their choices across periods. Early in learning (in Period 1) the most frequent choice was to study the pairs again and therefore not to attempt recall at the end of Period 1. The proportion of study choices decreased across periods,  $F(3, 87) = 53.20$ ,  $\eta_p^2 = .65$ , while the proportion of drop choices increased,  $F(3,$

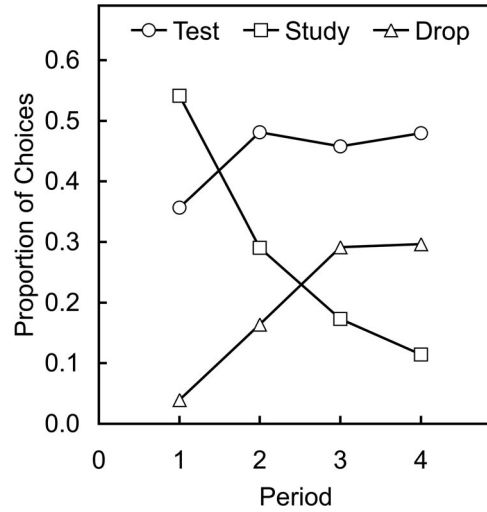


Figure 10. Proportion of Drop, Study, and Test choices in each period in Experiment 4.

87) = 6.49,  $\eta_p^2 = .18$ , and the proportion of test choices increased slightly,  $F(3, 87) = 2.49$ ,  $\eta_p^2 = .08$ . The result indicates that subjects did not choose to test themselves on the entire list early in learning—even though attempting recall would have improved their ability to encode items during the next study period.

**Strategy selection criteria.** The relationship between subjective experience and strategy selection was evaluated within the context of the schematic model described earlier. First, the relationships between JOLs and strategy choices were assessed as the mean within-subject gamma correlations between JOLs and the decision to drop or practice pairs and between JOLs and the decision to test or study them. The JOL-Drop gamma averaged .83 (one subject did not drop any pairs, and thus no correlation was calculated for this subject), and the JOL-Test gamma averaged .82. Therefore there was a strong correspondence between subjects' JOLs and their strategy choices.

The computational procedure used in Experiment 2 was used again to estimate the Drop and Test criteria ( $D_C$  and  $T_C$ , respectively) for each subject in the present experiment, with the candidate criteria that maximized each fit ratio considered the estimates of  $D_C$  and  $T_C$ . The mean  $D_C$  estimate was .86 (fit ratio = .91), and the mean  $T_C$  estimate was .43. (fit ratio = .84),  $F(1, 29) = 88.27$ ,  $\eta_p^2 = .75$ . Thus subjects set a high criterion for dropping pairs and a moderate criterion for testing pairs, just as they did in Experiment 2 (where the respective criteria were .88 and .56). The consequence of this control policy was that most students continued studying low-JOL items rather than testing them, despite the fact that attempting retrieval would have potentiated and enhanced learning.

Discussion

Experiment 4 showed that although retrieval facilitates encoding during restudy, subjects did not choose to attempt retrieval of all items and instead chose to continue studying items without testing. This strategy choice had negative consequences for learning. The analyses of metacognitive control showed that there was a strong

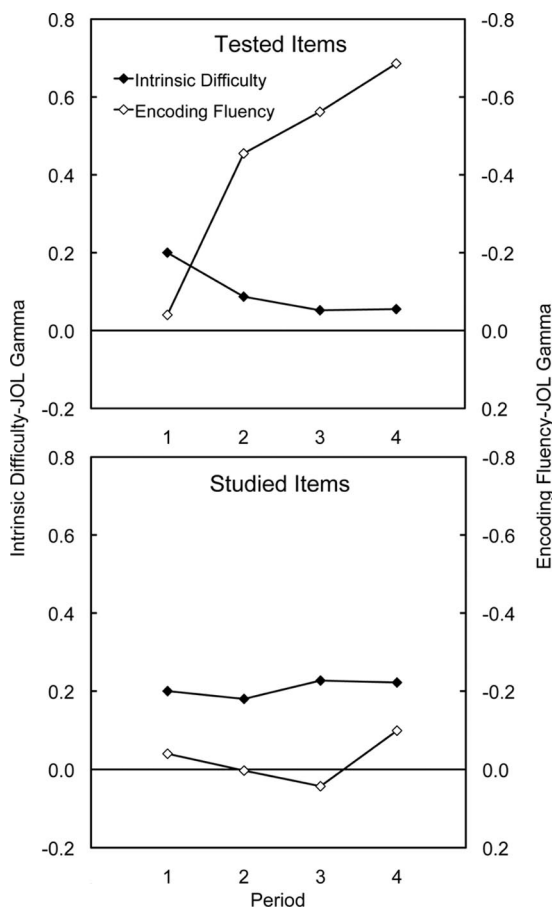


Figure 9. Gamma correlations between intrinsic difficulty (ease-of-learning ratings) and judgments of learning (JOLs; plotted on the left ordinate) and encoding fluency and JOLs (plotted on the right ordinate) in Experiment 4, showing a shift in the bases of JOLs for tested items but not repeatedly studied items.

correspondence between subjects' assessments of their own learning and their strategy choices. The results support the idea that subjects establish a criterion for dropping pairs and a criterion for attempting retrieval. However, the particular criteria that subjects established led them to continue studying pairs when JOLs were relatively low—even though attempting retrieval would have improved subjects' ability to encode and learn them.

### General Discussion

The goal of this project was to examine the interrelationships among three factors: (1) the effects of retrieval practice on learning, (2) students' monitoring of their own learning, and (3) students' decisions to practice retrieval in self-controlled learning conditions. In the following sections I present an overview of four main findings and discuss the implications of these findings.

#### *Overview of the Findings*

*The effects of retrieval on learning.* Practicing retrieval facilitated performance during the initial learning phase and on a delayed test 1 week later. Once a word pair could be recalled, practicing retrieval two additional times promoted long-term retention much more than studying it two additional times or removing it from further practice, and additional studying produced little benefit relative to dropping items. The results are consistent with prior research on repeated retrieval practice (e.g., Karpicke & Roediger, 2007, 2008) and show that the act of retrieval itself produces learning. In addition, inserting tests in the learning phase facilitated learning more than repeatedly studying without testing, a phenomenon known as the potentiating effect of retrieval (Izawa, 1969; Karpicke & Roediger, 2007). Attempting retrieval enhanced encoding in subsequent study periods even when the retrieval attempts failed.

*Metacognitive control and strategy selection.* When subjects were given control over their own learning they did not always choose to practice retrieval and performance suffered as a consequence. First, when subjects selected strategies during test periods they tended to drop pairs from further practice once they could recall them rather than practice retrieval. This produced poor long-term retention relative to what likely would have occurred with repeated retrieval practice. Second, when subjects selected strategies during study periods they often chose to continue studying pairs without attempting retrieval, especially early in the learning phase (cf. Kornell & Son, 2009). This strategy produced worse learning performance than attempting retrieval early in learning because retrieval attempts potentiate the learning that occurs in subsequent study periods. These two key results show that subjects do not always leverage retrieval practice to bolster their learning in self-regulated conditions.

*Metacognitive monitoring guides control.* Subjects' judgments of learning played an important role in guiding strategy choices, supporting the concept of a causal link from metacognitive monitoring to metacognitive control (Koriat & Goldsmith, 1996; Nelson & Narens, 1990). Subjects established two decision criteria to select strategies: a criterion for deciding to terminate or continue practice ( $D_C$ ) and, if subjects continued practice, a second criterion for deciding whether to test or continue studying ( $T_C$ ). Subjects evaluated their own learning, and most subjects based

their strategy choices on their JOLs. By and large, subjects established a control policy wherein their criterion for dropping pairs was relatively high (.88 and .86 in Experiments 2 and 4, respectively) and their criterion for testing pairs was near a moderate JOL level (.56 and .43 in Experiments 2 and 4, respectively). The decision criteria had different consequences in study and test periods. In test periods, when subjects correctly recalled an item their JOLs were often high—above the drop criterion—and therefore subjects dropped items instead of practicing retrieval. In study periods, when JOLs were low they fell below the test criterion and thus subjects continued studying items—despite the fact that attempting retrieval probably would have improved subsequent encoding even if the attempt failed.

*Multiple bases of judgments of learning.* JOLs were based on a variety of cues available to the learner, consistent with Koriat's (1997) cue utilization theory. The dominant cues for JOLs differed in study and test periods. When subjects made JOLs in test periods (in Experiments 1 and 2), JOLs were driven largely by the speed with which subjects could retrieve the target item (retrieval fluency), with quick retrieval leading to higher JOLs and slow retrieval leading to lower JOLs (cf. Benjamin et al., 1998). The pattern of results for JOLs in study periods was somewhat more complex. Early in learning, subjects tended to base JOLs on the intrinsic difficulty of items. However, as the learning phase progressed subjects shifted from reliance on intrinsic cues to reliance on internal mnemonic cues, specifically encoding fluency (cf. Koriat, 1997; Koriat et al., 2002). The present results showed that this shift in cues depended on test experience. Subjects exhibited this shift when they learned a list across alternating study and test periods but they continued to rely on intrinsic difficulty when they repeatedly studied without attempting retrieval.

#### *Implications for the Testing Effect*

There are implications of the current findings for theories of the testing effect. For instance, the present experiments examined several measures of objective and subjective item difficulty. The consistent result was that retrieval practice enhanced long-term retention regardless of item difficulty and regardless of how difficulty was assessed. Therefore it is not the case that mere selective practice of inherently easy items is responsible for the testing effect. On the contrary the present results show that retrieval practice promotes learning of both easy and difficult items, a finding with theoretical as well as practical implications for student learning.

The present research also provides a first step toward examining the potentiating effect of retrieval, which has received little attention since Izawa's research in the 1960s (see Karpicke & Roediger, 2007). The implication of the potentiating effect is that retrieval attempts enhance the processing that occurs in subsequent study and thereby improve learning. But the mechanisms responsible for the effect have not been examined in depth. Experiment 3 showed that the potentiating effect occurred for both easy and difficult items, so the effect is not simply due to improved learning of inherently easy items. The experiment also identified changes that occur in study periods following testing. Specifically, subjects increasingly relied on encoding fluency as a cue for JOLs as the learning phase progressed (Koriat, 1997), but reliance on encoding fluency depended on test experience. When subjects repeatedly

studied without testing they continued to base their JOLs to a large extent on intrinsic difficulty. Thus a fundamental change in the information used by subjects to assess their own learning may be partly responsible for the potentiating effect of retrieval.

### *Implications for Metacognitive Monitoring*

The present experiments were inspired in part by the cue-utilization framework of Koriat (1997), and indeed the majority of the results can be interpreted within that framework. A key factor in the present experiments was whether JOLs were obtained during test periods (Experiments 1 and 2) or study periods (Experiments 3 and 4). Different cues underlie JOLs depending on when the JOLs occur. JOLs during test periods were governed largely by retrieval fluency (cf. Benjamin et al., 1998) whereas JOLs during study periods were influenced more by encoding fluency (cf. Koriat & Ma'ayan, 2005). Further, while subjects' JOLs were often sensitive to the perceived intrinsic difficulty of items, they were often not sensitive to other potential cues like the number of trials it took to recall an item for the first time (cf. Koriat, 2008). The result is consistent with the theory that JOLs are comparative in nature and tuned more to relative differences among items than to factors associated with overall levels of performance (cf. Koriat, 1997; Koriat et al., 2006). The results of Experiments 3 and 4 also showed that the shift from intrinsic cues to internal mnemonic cues for JOLs depends to a large extent on testing during learning.

### *Toward a Broader Conceptualization of Metacognitive Control*

There has been a wealth of research on study-time allocation (Dunlosky & Thiede, 1998; Metcalfe & Kornell, 2003; Nelson & Leonesio, 1988; Son & Metcalfe, 2000; Thiede & Dunlosky, 1999). The assumption guiding this prior research has been that improving the amount of study time allocated to individual items is critical to promoting learning. Some authors have made this assumption explicit (see Nelson & Leonesio, 1988; Nelson & Narens, 1990) and it is implicit in many studies of self-paced study time. But there are many reasons to question this assumption. For example, spending additional time studying sometimes produces no learning, a result found with simple paired associates (Nelson & Leonesio, 1988) as well as educational text materials (Callender & McDaniel, 2009). Study time and learning can be negatively correlated—conditions that involve spending more time studying items can produce less learning than conditions requiring less study time (Craig & Tulving, 1975). The fact is that study time itself does not cause learning—it is processes occurring over time that cause learning. And when students regulate and control their own learning it is the encoding processes that students select that cause learning or fail to produce it. The relationship between metacognitive control and learning is not simply an issue of how much study time subjects allocate to particular items—it is an issue of what strategies subjects recruit to learn particular materials in a given context.

The time is ripe to examine how students select and implement learning strategies, not just how they allocate study time. The present research is just one example of how that might be accomplished and it shows that the control policy that students adopt has

important implications for the strategies they choose and consequent learning.

### *Practical Implications for Self-Regulated Learning*

This research on the metacognitive processes involved in strategy selection has identified a robust illusion that occurs during self-regulated learning: When students study on their own they often base strategy choices on their judgments of learning, yet these judgments may lead students to select poor strategies due to the control policies that guide their choices. For instance, consider a student in Experiment 2 who correctly recalled a word pair very easily and assigned it a 100% JOL rating because of his or her fluent retrieval. As a consequence of the high JOL—and the placement of his or her criterion for dropping items ( $D_c$ )—the student would likely decide to remove the pair from further practice even though the probability that he or she would recall the pair on the final test would likely be less than 100% (given that average recall of dropped items was about 35% in Experiments 1 and 2). The strategies that students select when they regulate their own learning are closely tied to subjective assessments of current performance, even though such assessments may not correspond to actual long-term future performance (cf. Bjork, 1999).

The results of Experiments 2 and 4 show that under certain circumstances students do choose to test themselves. Kornell and Son (2009) also recently showed that students sometimes choose to test themselves—but like the results of the present experiments, they found that the reason students test themselves is often to find out how well they have learned an item. That is, if students test themselves they likely do it to diagnose their knowledge rather than to practice retrieval per se. The results of Experiment 4 also show that students often wait until they have reached a sufficient JOL level before they attempt retrieval, even though testing earlier in learning would have potentiated and improved learning. Though some students do test themselves while they are learning they may not always implement self-testing in the most effective way to promote the best learning and long-term retention.

Research on the monitoring and control processes involved in strategy selection has the potential to inform practices that may help students recruit more effective learning strategies. One way to improve self-regulated learning could be to improve the accuracy of metacognitive monitoring by debiasing subjects' predictions. For example, in Experiments 1 and 2 subjects' JOLs were generally overconfident relative to actual performance on delayed tests. Perhaps this overconfidence is due in part to subjects' insensitivity to the effects of the 1-week retention interval on performance. Therefore activating the notion of forgetting in the minds of students may improve the accuracy of their JOLs (cf. Koriat et al., 2004). But even if the accuracy of metacognitive judgments were improved, students must still possess a theory about why particular strategies are effective. That is, even if students make accurate JOLs they still might not practice repeated retrieval because they lack a theory about why retrieval practice promotes retention (though Experiment 2 showed that some students do possess such a theory). Indeed, survey data show that a majority of college students seem to base their study strategies on the theory that repeated reading promotes learning and do not always see the value of practicing retrieval (Karpicke, Butler, & Roediger, 2009). Changing the control policy adopted by subjects may be more



important for promoting learning than improving the accuracy of metacognitive monitoring. These applied questions about how to help students choose effective learning strategies can be informed by basic research like the present experiments.

### Conclusion

Retrieval practice is a potent way to improve learning and long-term retention. However, when students regulate their own learning, by and large they do not choose to practice retrieval. Once they can recall a fact they believe they know it and often choose to drop it from practice. Students also tend to wait until they have reached a level of confidence in their learning before they begin to practice retrieval, even though attempting retrieval would potentiate and improve encoding the next time they studied even if the retrieval attempt failed. Both findings demonstrate illusions that occur during self-regulated learning. The strategies that students select are closely tied to their assessments of their own learning and self-selected strategies have significant consequences for learning. Despite the fact that most memory research has traditionally exercised strict experimental control over the mnemonic activities of the learner, the strategies that people select on their own are worth examining in their own right.

### References

- 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. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.
- 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, NY: Wiley.
- 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.
- Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*, 30–41.
- Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative processing explanation of the testing effect. *Memory & Cognition*, *34*, 268–276.
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.
- Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.
- Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1180–1191.
- Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica*, *98*, 37–56.
- Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.
- Izawa, C. (1969). Comparison of reinforcement and test trials in paired-associate learning. *Journal of Experimental Psychology*, *81*, 600–603.
- Izawa, C. (1971). The test-trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224.
- Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*, *17*, 471–479.
- Karpicke, J. D., & Roediger, H. L. (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, February 15). The critical importance of retrieval for learning. *Science*, *319*, 966–968.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.
- Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 289–325). New York, NY: Cambridge University Press.
- Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*, 416–428.
- Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643–656.
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.
- Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*, 478–492.
- 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.
- Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*, 147–162.
- 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. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.
- Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501.
- Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 464–470.
- Metcalf, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*, 159–163.
- Metcalf, 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.
- Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*, 325–335.
- Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 676–686.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–141). New York, NY: Academic Press.
- 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.
- Pressley, M., & Harris, K. R. (2006). Cognitive strategies instruction: From

- basic research to classroom instruction. In P. A. Alexander & P. Winne (Eds.), *Handbook of educational psychology*. New York, NY: MacMillan.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447.
- Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
- Roediger, H. L., & Karpicke, J. D. (2006b). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.
- 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.
- Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review*, *71*, 219–237.

## Appendix

### Results of Separate 2 (Measure) $\times$ 4 (Period) Analyses of Variance for the Drop, Study, and Test Conditions in Experiment 1

Condition	<i>dfs</i>	<i>F</i>	$\eta_p^2$
<b>Drop</b>			
Main effect of measure	1, 396	110.77	.22
Main effect of period	3, 396	0.81	<i>ns</i>
Interaction	3, 396	3.13	.02
<b>Study</b>			
Main effect of measure	1, 396	378.43	.49
Main effect of period	3, 396	6.19	.05
Interaction	3, 396	10.53	.07
<b>Test</b>			
Main effect of measure	1, 396	6.24	.02
Main effect of period	3, 396	3.78	.03
Interaction	3, 396	0.88	<i>ns</i>

*Note.* Measure = judgment of learning versus final recall; period = period in which items were first recalled.

Received August 7, 2008  
 Revision received July 8, 2009  
 Accepted July 9, 2009 ■