Journal of Experimental Psychology: Applied
2013, Vol. 19, No. 1, 1–13

1076-898X/13/$12.00 DOI: 10.1037/a0032147

Feedback Reduces the Metacognitive Benefit of Tests

Nate Kornell
Williams College

Matthew G. Rhodes
Colorado State University

Testing long-term memory has dual benefits: It enhances learning and it helps learners discriminate what they know from what they do not know. The latter benefit, known as delayed judgment of learning (dJOL) effect, has been well documented, but in prior research participants have not been provided with test feedback. Yet when people study they almost universally (a) get feedback and (b) judge their learning subsequent to receiving the feedback. Thus, in the first three experiments, participants made JOLs following tests with feedback. Adding feedback significantly decreased the dJOL effect relative to conditions taking a test without receiving feedback. In Experiment 4, participants made decisions about which items to restudy (without actually restudying); adding feedback also decreased the accuracy of these decisions. These findings suggest that, in realistic situations, tests enhance self-monitoring, but not as much as previously thought. Judging memory based on prior test performance and ignoring the effects of feedback appears to produce an "illusion of not knowing."

Keywords: learning, metacognition, feedback, test, judgment of learning

Retrieving information from memory has at least two important benefits. One is that taking a test enhances learning, compared with studying without taking a test (i.e., the *testing effect*; see Roediger & Butler, 2011; Roediger & Karpicke, 2006b, for reviews). The focus of the current article, however, is on the metacognitive benefit of retrieval: Taking a test allows people to accurately monitor their learning and distinguish what they know from what they do not know (e.g., King, Zechmeister, & Shaughnessy, 1980; cf. Rhodes & Tauber, 2011a). The majority of students who test themselves while studying do so for the second reason, to figure out what they do and do not know (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Kornell & Son, 2009).

The metacognitive value of tests was demonstrated in seminal research on the delayed judgment of learning (dJOL) effect (Dunlosky & Nelson, 1992; see also Nelson & Dunlosky, 1991). Participants first studied a set of unrelated word pairs (e.g., *ocean-tree*) and then made a judgment of learning (JOL) about each pair, indicating the probability that they would remember the target when shown the cue on a later test. In some cases, the JOL was made immediately after the pair was studied and in other cases the JOL was made after a delay. This manipulation of JOL timing was crossed with a manipulation of the nature of the cue used to elicit the JOL. Half of the participants were shown the cue and target (e.g., *ocean-tree*) prior to the JOL prompt; the other half of the participants were shown only the cue (e.g., *ocean-?*). JOLs were then assessed with respect to the degree to which they distinguished between what was or was not remembered (i.e., *relative accuracy*), operationalized via the correlation between JOLs and final test performance. This correlation was greater for delayed cue-only JOLs (gamma correlation = .93) than in the other three conditions (gammas < = .60). Subsequent research has replicated this finding many times (see Rhodes & Tauber, 2011a, for a meta-analytic review). In short, the condition that produced the greatest correlations between JOLs and recall—the delayed cue-only condition—was the one that allowed for a meaningful test of recall from long-term memory.

In this article, we highlight two shortcomings of previous research on the dJOL effect. Both shortcomings derive from the fact that participants were not provided with correct answers (i.e., feedback) before making JOLs in any previous studies. (As we discuss shortly, Kimball & Metcalfe, 2003, provided feedback *after* participants made JOLs, and Rhodes & Tauber, 2011b, provided feedback, but not the correct answer, before participants made JOLs). In the absence of feedback it is difficult to know *why* there is a dJOL effect, a theoretical issue to which we return shortly. First, we discuss a second, more practical problem: When students test themselves in the course of studying, they customarily check the answer soon afterward; that is, tests without feedback are not a common study technique.

How Do Students Actually Study?

Students study in many ways. Some test themselves and others do not (see, e.g., Hartwig & Dunlosky, 2012), and of course the content being studied is important; self-testing using flashcards is common in organic chemistry and first-year Spanish, but not in American literature. Such self-testing is not limited to flashcards. For example, a student studying a textbook might take the test at the end of the chapter and then check the answers. In both cases, the student will make JOLs (often implicitly) about their accuracy, which often translates into decisions about what to study (Nelson,

Nate Kornell, Department of Psychology, Williams College; Matthew G. Rhodes, Department of Psychology, Colorado State University.

Katie Flanagan deserves special thanks for her assistance in collecting and analyzing the data from these experiments.

Correspondence concerning this article should be addressed to Nate Kornell, 18 Hoxsey Street, Williamstown, MA 01267. E-mail: nkornell@gmail.com

1996; Nelson & Narens, 1990, 1994; but see also Koriat, Ma'ayan, & Nussinson, 2006).

Although students study in many ways, there are also ways they rarely study. One is to test oneself and then not check the answer afterward. Yet this is exactly what occurs in the test condition of almost all previous research on the dJOL effect (the main exception is Kimball & Metcalfe, 2003, but as we discuss later, this study was unrealistic in its own way). The dJOL effect has not yet been established using a paradigm that reflects the way students actually study.

Many researchers recommend that students test themselves to enhance their metacognitive monitoring (including us, e.g., Kornell & Son, 2009; Rhodes & Tauber, 2011a). The idea is that self-testing will improve their study decisions and, ultimately, their learning. Such recommendations are premature without examining the dJOL effect in a realistic paradigm. Thus, we examined the metacognitive value of tests using a paradigm in which participants took a test, received feedback, and then made JOLs.

There is reason to predict that tests might not enhance monitoring as much as previously thought. People tend to make JOLs based on their memory of a prior test (MPT; Finn & Metcalfe, 2007, 2008; but see also Ariel & Dunlosky, 2011; Tauber & Rhodes, 2012). In relying on this MPT heuristic, they underestimate the effect that subsequent learning—from feedback—has on their knowledge. Thus, as a result of feedback, participants might learn items without being aware that they have learned those items, potentially decreasing metacognitive accuracy.

Prior research has also established, however, that making answers available might make metacognitive judgments more accurate. Students are often overconfident in the accuracy of their recall of definitions from a text passage. Providing them with the correct answers seems to ameliorate this overconfidence, in particular for incorrect responses (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Lipko et al., 2009; Rawson & Dunlosky, 2007). However, these findings concerned retrospective confidence judgments about the accuracy of a prior response, with analyses focused on mean differences between recall and confidence. In contrast, JOLs are predictions of future remembering and our primary interest was the correlation between confidence and accuracy. Because of these differences, we predicted that, when making JOLs, providing the answer would decrease metacognitive accuracy.

Causes of the Delayed-JOL Effect

The second motivation for the experiments we report was theoretical. One theoretical issue, mentioned above, was testing the prediction that participants would ignore feedback, even if it immediately preceded a JOL, and base their judgments on their response to a prior test. A second theoretical issue had to do with the cause of the dJOL effect.

Feedback may help solve a problem first pointed out by Spellman and Bjork (1992): There is more than one way to explain the finding that tests enhance JOL accuracy. They distinguished between the metamemory hypothesis and the memory hypothesis. According to the *metamemory hypothesis*, tests of long-term memory are diagnostic of later recall; knowing what one can and cannot recall leads to accurate JOLs. For example, the monitoring dual memories account of the delayed JOL effect (Dunlosky & Nelson, 1992, 1994; Nelson & Dunlosky, 1991) suggests that immediate JOLs are only moderately accurate because they largely rely on information derived from short-term memory (STM). Such information is noisy and thus only weakly related to future memory performance. Conversely, a delay shifts the basis of JOLs to long-term memory, making them more accurate.

According to the *memory hypothesis*, though, tests are not just diagnostic of later recall—they actually change later recall in a way that makes delayed JOLs seem more accurate than immediate JOLs (Spellman & Bjork, 1992; see also Spellman, Bloomfield, & Bjork, 2008). If an item is tested, it is either recalled or not. If it is not recalled and feedback is not provided, there is little chance the item will be recalled on the final test, and the item generally receives a low JOL. If the item is recalled, however, it becomes stronger as a result of the test trial (i.e., more likely to be recalled on a subsequent test) and is generally given a high JOL. The net result is that tests without feedback strengthen strong items without strengthening weak items, separating items into two relatively distinct distributions: Strong items that are very well-learned and weak items that are very unlikely to be recalled later (Kornell, Bjork, & Garcia, 2011). According to the memory hypothesis, even if monitoring accuracy were not affected by tests, taking the test changes the memory to conform to the judgment (i.e., items given high JOLs become stronger and items given low JOLs do not)—making monitoring accuracy appear to increase because of tests.

Providing feedback may help ameliorate this problem, because even items that are not retrieved can benefit from test/feedback trials (e.g., Grimaldi & Karpicke, 2012; Karpicke, 2009; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009). This across-the-board benefit prevents the bifurcation of item distributions (Kornell et al., 2011). Accordingly, the memory hypothesis predicts that tests followed by feedback should lead to little or no increase in metacognitive accuracy, compared with trials without tests. The metamemory hypothesis seems to make a different prediction: Tests should enhance JOL accuracy, compared with trials without tests, even if feedback is provided (Kimball & Metcalfe, 2003). However, if people undervalue feedback when making JOLs, and instead rely on the MPT heuristic, then JOL accuracy might be lower for tests with feedback than for tests without feedback.

Kimball and Metcalfe (2003) attempted to contrast the memory and metamemory hypotheses by providing feedback in a delayed JOL study. Unlike the experiments we report, participants took a test, made a JOL, and then received feedback. Under these conditions, the delayed-JOL effect disappeared, supporting the memory hypothesis over the metamemory hypothesis. However, this finding is difficult to interpret because participants received feedback *after* making JOLs. It seems likely that participants would have changed their JOLs following the feedback on some trials (e.g., when they thought they had been right but found out they were wrong or vice versa), but such changes were impossible. In addition, providing feedback changed participants' memory state after they had judged that memory state. Thus, this procedure may have artifactually diminished the correlation between judgments and eventual memory performance. The procedure was also problematic from a practical perspective because it prevented participants from updating their JOLs after being shown an answer,

whereas learners are usually free to adjust their self-monitoring in an ongoing fashion.[1]

In the present studies we gave participants a test, provided feedback in the form of the correct answer, and then solicited a JOL. Thus, in contrast to Kimball and Metcalfe's (2003) study, our participants had full information when making their judgments and faced a paradigm that corresponded to real studying. Our underlying logic was the same as theirs: According to the metamemory hypothesis, a dJOL effect should remain in the presence of feedback (although feedback might weaken it); according to the memory hypothesis, preventing the bifurcation of recalled and unrecalled items should largely eliminate the dJOL effect.

No prior research has asked a question, provided the correct answer, and then asked for a JOL, but in study by Weaver and Kelemen (2003), participants were shown a cue alone and were presented with the correct answer below it, distributed among other correct answers. Thus, it was possible for participants to test themselves before seeking out the correct answer if they wanted to, but they could also check the correct answer immediately. Conditions in which the correct answer was available led to less accurate metacognitive judgments. If participants were indeed testing themselves before seeing the feedback, this finding would predict that feedback should produce a similar decrement in metacognitive accuracy, but the extent to which they did so is not clear.

We know of only one study in which participants took a test, and then received feedback, prior to making a JOL. Rhodes and Tauber (2011b) used deceptive items that enticed participants to provide an incorrect answer (e.g., for the pair table-ch__r, participants frequently recalled the incorrect semantically related competitor "chair" rather than the target "cheer") prior to making an immediate or delayed JOL. They found that deceptive items eradicated the delayed-JOL effect, but they also observed that providing feedback on a pre-JOL recall attempt reinstated the delayed JOL effect (Experiment 3). However, the feedback only indicated whether the participant had answered correctly or not and did not provide the correct answer. In the research we report here, we provided the correct answer as feedback. Providing the correct answer is important for our purposes because it prevents bifurcation of recalled versus unrecalled items. It also corresponds to how people prefer to study.

The Present Research

We compared the same three conditions in four experiments. After an initial study phase, participants in the Read-Only condition were shown a cue and target then made a JOL; participants in the Test-Only condition were shown a cue alone and made a JOL; participants in the Test-With-Feedback condition were shown a cue alone, tried to think of the target, and were then given feedback, after which they made a JOL. In all conditions, a final test phase followed the JOL phase. In Experiment 4, the JOLs were replaced with study choices.

Experiments 1–3 were increasingly realistic attempts to examine how tests affect the accuracy of JOLs. Experiment 1 was a relatively standard JOL paradigm, except that, like most real studying, all timing was under the participant's control. In Experiment 2, participants were not asked to type in their answers (to mimic real online and paper flashcard use). The same was true in Experiment 3 and, in addition, the learning materials were changed from paired associates (e.g., *thunder-noise*) to a novel episodic memory task, learning Indonesian foreign language vocabulary (e.g., *kiss-ciuman*). Because JOLs play an important role in study decisions, participants in Experiment 4 were asked to decide which items they did (or did not) want to restudy in a procedure modeled on Experiment 3.

Experiment 1

One of our primary goals was to explore the dJOL effect in a realistic paradigm. An informal investigation of popular online flashcard programs (e.g., Anki, StudyBlue) revealed that users control how much time they spend studying each item. This is obviously true with paper flashcards—and almost all other forms of studying outside of class—as well. Thus, our participants controlled the presentation duration as they studied.

In Experiment 1, participants studied the 36 word pairs, then did a second phase with three conditions: They restudied, took a test without feedback, or took a test with feedback. Unlike Experiments 2–4, they were asked to type in the target during the second phase. Immediately after each trial they made a JOL. After a short distractor task, participants took a test on all items.

Method

Participants. One hundred forty-six participants (86 women, 60 men; median age = 28 years, range = 19–74 years) were paid $1.00 for completing the experiment, which took 15–20 min. They were recruited via Amazon's Mechanical Turk, a Web site that allows users to complete small tasks for pay. A number of recent studies have shown that Mechanical Turk produces the same findings as laboratory-based methods of data collection (Buhrmester, Kwang, & Gosling, 2011; Kittur, Chi, & Suh, 2008; Mason & Suri, 2012; Sprouse, 2011). Recruitment was limited to participants living in the United States, and participants could not participate in more than one of the studies reported in this article. Because of random assignment, there were 47, 47, and 52 participants in the Test-Only, Test-With-Feedback, and Read-Only conditions.

Materials. The materials consisted of 36 weakly related word pairs (e.g., *abdomen-organ; jack-hammer*). All pairs had forward association strengths between .02 and .04 (Nelson, McEvoy, & Schreiber, 1998). In other words, when shown the cue, 2%–4% of people produced the target as their first response.

Procedure. The procedure comprised three stages. After reading instructions explaining the procedure, participants studied 36 word pairs. On each trial, a cue and target were presented together (e.g., *nerve-center*), and the participant pressed a button to move on to the next pair.

The second stage was the judgment phase. There were three between-participants conditions. Prior to the JOL prompt, in the Read-Only condition, the cue and target were presented together

[1] The idea that participants change their JOLs based on feedback may seem inconsistent with the MPT heuristic. We believe that participants ignore feedback when it does not change their judgments (i.e., when they thought they got the answer wrong and find out that they did indeed get it wrong). Their judgments may be influenced by feedback, however, when their assessment was wrong; that is, when they thought they were wrong but find out they were actually right, or vice versa.

(e.g., *nerve-center*) until participants pressed a button onscreen. In the Test-Only condition, only the cue was presented (e.g., *nerve*); participants were asked to type in the target and then press a button onscreen. In the Test-With-Feedback condition, the cue was presented (e.g., *nerve*) and participants were asked to type in the target and then press a button onscreen. When they did, the same cue was presented with the target (e.g., *nerve-center*) until the participant pressed the button onscreen to move on. At this point, in all cases, participants were shown a JOL prompt that read "Chance you'll recall the answer later (0–100)" and typed in a JOL. Next, they pressed a button to move on to the next trial.

After the second phase, participants played the video game Tetris for 3 min. They were then given a test, during which each cue was shown, one at a time. Participants were asked to type in the target and then press enter.

Data Analysis

Answers were scored as correct if they were accurate or close to accurate. Any response that scored 75 or greater using the similar_text function in the programming language PHP was scored as correct.

Consistent with prior work on the dJOL effect (e.g., Nelson & Dunlosky, 1991), the key analysis focused on the relative accuracy of JOLs, operationalized via the gamma correlation (Nelson, 1984). This nonparametric measure of association quantifies the degree to which a JOL for an individual item predicts later memory accuracy on that item. Gamma correlations will be positive when subsequently remembered items are given high JOLs and items that are not remembered are given lower JOLs. Gamma correlations were computed individually for each participant and then averaged.

Results

Data from four participants were excluded from the analyses because their gamma correlations were more than 2.5 *SD*s below the mean gamma correlation (two from the Read-Only condition and one from each of the other conditions). Another 13 participants were excluded for whom gamma correlations could not be computed, either because their test accuracy was perfect or because they gave the same JOL for every item. Thus, gamma correlations were analyzed for 43, 46, and 40 participants in the Test-Only, Read-Only, and Test-With-Feedback conditions, respectively. No participants were excluded from the nongamma analyses.

On the initial test, during the study phase, participants in the Test-Only and Test-With-Feedback conditions answered correctly on 58% and 56% of trials, respectively (*SD* = 26% and 24%, respectively). This difference was not significant, $t(92) = .37$, $p = .71$. This finding contradicts the possible concern that participants could have invested less effort in the retrieval task during the study phase when they were anticipating imminent feedback.

JOL magnitude, final cued-recall accuracy, and correlations between recall accuracy and JOLs (i.e., resolution) are presented in Table 1. Gammas correlations were significantly affected by study condition, $F(2, 126) = 41.98$, $p < .0001$, $\eta_p^2 = .40$. As predicted, the correlation was highest in the Test-Only condition and lowest in the Read-Only condition; the Test-With-Feedback condition fell in between. Tukey-Kramer post hoc analyses confirmed that all three conditions differed significantly from each other (mean difference = 30.3, critical difference = 14.6 for Test-Only vs. Test-With-Feedback; mean difference = 54.3, critical difference = 14.1 for Test-Only vs. Read-Only; mean difference = 24.0, critical difference = 14.4 for Test-With-Feedback vs. Read-Only).

Table 1
Mean Correlations, Judgments of Learning (JOLs), and Percentage Correct on the Final Test in Experiment 1

	Gamma	JOL	Accuracy
Test only	.851 (.173)	60.1 (26.7)	56.9 (26.5)
Test with feedback	.548 (.285)	60.2 (21.3)	82.4 (14.5)
Read only	.308 (.347)	62.2 (17.1)	78.1 (18.1)

Note. *SD*s are in parentheses.

JOL magnitude was not significantly influenced by study condition, $F(2, 143) = .15$, $p = .86$. Final recall accuracy was significantly affected by study condition, $F(2, 143) = 21.61$, $p < .0001$, $\eta_p^2 = .23$. As Table 1 shows, low accuracy in the Test-Only condition largely drove this effect.

Previous research would seem to predict that recall should be higher in the Test-With-Feedback condition than in the Read-Only condition. Thus, we conducted a one-tailed planned comparison between these two conditions. Though very weak, the testing effect was marginally significant $t(97) = 1.30$, $p = .099$. The weakness of this effect is consistent with recent evidence showing that making JOLs can diminish testing effects, as we discuss in the General Discussion.

Were JOLs affected by feedback? One possible explanation of why feedback made metacognitive judgments less accurate is that participants failed to adequately account for learning that occurred as a result of feedback. Feedback clearly affected learning, but did it affect JOLs? To investigate this question we compared JOLs in the Test-Only condition and the Test-With-Feedback condition.

The Test-Only and the Test-With-Feedback conditions differed drastically in recall accuracy (56.9% and 82.4%, respectively). Mean JOLs did not differ (60.1% and 60.2%, respectively). This analysis seems to suggest that feedback had little effect on JOLs, but it does not tell the whole story. During the study phase, if a participant made a correct response, feedback would not be expected to affect their JOLs because feedback has little or no effect on learning following correct responses (e.g., Hays, Kornell, & Bjork, 2010; Pashler, Cepeda, Wixted, & Rohrer, 2005). It is only following incorrect responses that feedback affected learning, and should therefore affect both JOLs and final test performance.

Table 2 displays average JOLs and final test accuracy analyzed separately for items answered correctly and incorrectly during the study phase (three participants were excluded because they made no errors). As expected, feedback following correct responses did not have a major impact on JOLs or final test accuracy. Responding following errors was more interesting. Feedback increased JOLs by 5.6 percentage points but it increased accuracy by 64.8 percentage points. The former difference was not significant, $t(89) = 1.0$, $p = .32$; the latter was significant, $t(89) = 17.9$, $p < .0001$. Thus, participants may have taken feedback into account, but they underestimated its importance.

Table 2
Mean Judgments of Learning (JOLs) and Percentage Correct on the Final Test in Experiment 1, Analyzed Separately Based on Whether the Corresponding Response During the Study Phase Was Correct or Incorrect

	JOL following error	JOL following correct	Accuracy following error	Accuracy following correct
Test with feedback	37.4 (22.9)	79.8 (19.7)	70.1 (21.1)	92.7 (10.9)
Test only	31.8 (30.3)	77.9 (22.5)	5.3 (12.2)	92.8 (10.6)
Difference	5.6	1.9	64.8	−0.1

Note. *SD*s are in parentheses.

Table 3
Mean Judgments of Learning (JOLs) in Experiment 1, Analyzed Separately Based on the Accuracy of the Corresponding Response During the Study Phase and Test Phase

	Test only		Test with feedback	
	T1 wrong	T1 correct	T1 wrong	T1 correct
Final test wrong	32.5	57.7	31.6	69.4
Final test correct	42.6	80.7	37.2	80.8
Difference	10.1	23.0	5.6	11.4

Note. Because of the small *N* in some cases, means were computed collapsed across participants rather than for each participant separately.

Averages can conceal differences in JOL distributions. Figure 1 displays the distribution of JOLs following correct and incorrect responses in the Test-Only and the Test-With-Feedback conditions. It appears that feedback did not affect JOLs following correct responses. It did, however, decrease the likelihood of very low JOLs following errors. Thus, again, it appears that feedback may have had a small effect JOLs. It is also clear that participants did not adequately account for feedback in their judgments; doing so would have entailed making more high JOLs when feedback was provided than when it was not, and the participants in Experiment 1 did not do so. If anything, they made fewer.

Did prior test performance control JOLs? Prior test performance influences JOLs (e.g., Finn & Metcalfe, 2008). We tested the strength of this influence by analyzing JOLs split into categories based on both initial and final recall accuracy. If memory for past test guides judgments, initial test performance should have a large influence on JOLs. As Table 3 shows, this hypothesis was supported. In fact, in both conditions, participants gave higher JOLs to items they got right initially but wrong on the final test than they did to items they got wrong initially but right on the final test—despite the fact that they were asked to judge final test performance. To compute *t* tests we analyzed data from 37 participants who had observations in each of the four possible cells of Table 3. The difference between JOLs for items that were incorrect initially but correct later versus correct initially but incorrect later was significant in the Test-Only condition (*M*s = 38.1 and 65.9, respectively), $t(11) = 2.6$, $p < .05$, and the Test-With-Feedback condition (*M*s = 39.1 and 70.0, respectively), $t(24) = 6.8$, $p < .0001$. Thus, prior test performance had a powerful effect on JOLs, even when feedback intervened between the test and the JOL.

Table 3 also suggests, however, that prior test performance was not the only influence on JOLs. If it had been, then JOLs following an error should have been the same regardless of whether the item was answered correctly or not on the final test. The same should be true of JOLs following a correct response. This hypothesis was not supported. In all four columns of Table 3, JOLs were higher for items that would subsequently be answered correctly than those that would not. Analyzing the 37 participants with observations in all cells showed that these differences were significant in the Test-With-Feedback condition, following errors, $t(24) = 3.37$, $p < .01$; following corrects, $t(24) = 2.52$, $p < .05$, but not in the

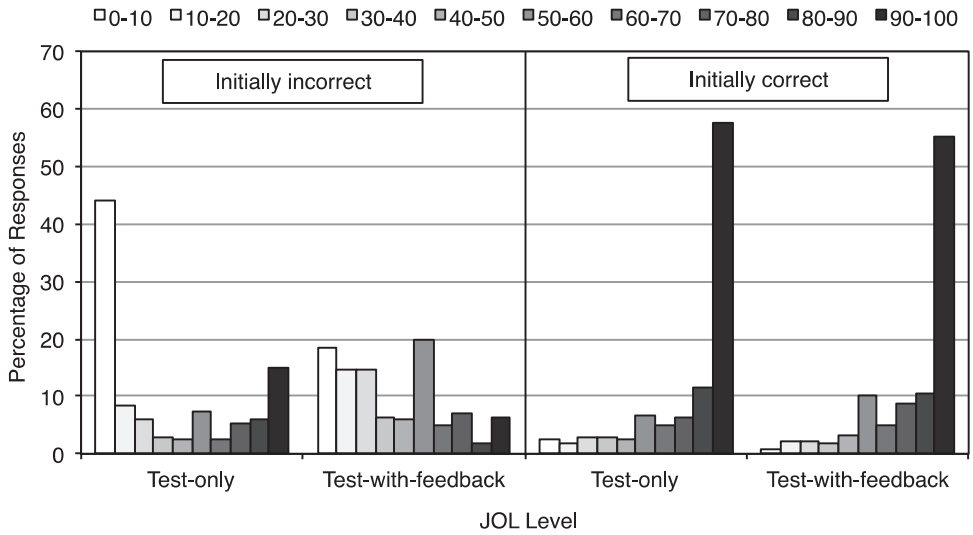

Figure 1. Percentage of judgment of learning (JOL) responses in the 0–10, 10–20, and so on, ranges in Experiment 1. Bars on the left side of the panel represent JOLs following an incorrect response; initially correct responses are on the right.

Test-Only condition, following errors, $t(11) = 1.40$, $p = .19$; following corrects, $t(11) = 1.54$, $p = .15$. This finding suggests that participants tapped into some cue or cues that went beyond prior test performance (similar conclusions can be found in Ariel & Dunlosky, 2011; Tauber & Rhodes, 2012). These data do not identify the nature of these cues. To summarize the findings from Table 3, memory of a prior test strongly influenced JOLs, but it was not the only influence.

Discussion

Tests enhanced JOL accuracy (i.e., the correlation between JOLs and recall accuracy), replicating previous findings (e.g., King et al., 1980). However, this enhancement in metacognitive accuracy diminished significantly when the tests were followed by feedback.

One possible explanation for this finding is that participants a) based their JOLs on their previous test and b) failed to take the benefit of feedback into account (see Finn & Metcalfe, 2008). The findings support both of these claims. Prior test performance strongly influenced JOLs (though it was not the only influence). Moreover, following errors, feedback had a much larger effect on recall than it did on JOLs. Feedback following an error does have a small effect on judgments, but it is dwarfed by the effect of feedback on recall.

Prior studies have shown that people can develop an "illusion of knowing" when told a correct answer (e.g., Glenberg, Wilkinson, & Epstein, 1982; Koriat, 1998). The present results suggest that after taking a test, receiving feedback produced an illusion of *not* knowing. That is, when participants could not recall answer and then received feedback, they usually judged that they would not know the answer on the final test either—a judgment that was frequently mistaken.

Experiment 2

The primary goal of Experiment 2 was to replicate Experiment 1 using a more realistic procedure. Flashcards, and most digital flashcard programs, do not require users to give overt answers while they learn; there is no requirement that anything be typed in or spoken aloud. Simply producing an answer overtly, even if it is not retrieved from memory, can affect learning (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Production also affects metacognitive judgments, both when production impacts recall and when it does not (Castel, Rhodes, & Friedman, 2013). Overt responding seemed like it could be an important potential difference between the research literature and actual practice. Thus, to increase realism, participants in Experiment 2 only typed in answers during the final test. Otherwise, the procedure was identical to the procedure in Experiment 1, permitting us to test the reliability of those findings.

Method

Participants. Ninety-two participants (61 women, 31 men; median age = 28 years, range = 18–66 years) were paid $1.00 for completing the experiment, which took 15–20 min. They were recruited via Amazon's Mechanical Turk. All participants reported living in the United States. Because of random assignment, there were 28, 32, and 32 participants in the Test-Only, Test-With-Feedback, and Read-Only conditions, respectively.

Procedure. The procedure was identical to the procedure in Experiment 1 with one exception. During the judgment phase, in the Test-Only and Test-With-Feedback conditions, participants were shown a cue and asked to think of the target. Unlike Experiment 1, though, they were not asked to type in an answer. After making a covert retrieval attempt they pressed a button to move on (they then made a JOL in the Test-Only condition or viewed the answer and then made a JOL in the Test-With-Feedback condition). The Read-Only condition was identical to Experiment 1.

Results and Discussion

When analyzing gamma correlations, data from nine participants were excluded (these participants were included in the other analyses), one (from the Test-With-Feedback condition) because his or her gamma correlation was more than 2.5 *SD* below the mean and another eight for whom gamma correlations could not be computed, either because their test accuracy was perfect or because they gave the same JOLs for every item. Thus, gamma correlations were analyzed for 27, 28, and 28 participants in the Test-Only, Read-Only, and Test-With-Feedback conditions, respectively.

JOL magnitude, recall accuracy, and correlations between recall accuracy and JOLs (i.e., resolution) are presented in Table 4. Gammas correlations were significantly affected by study condition, $F(2, 80) = 34.44$, $p < .0001$, $\eta_p^2 = .46$. Consistent with Experiment 1, Tukey-Kramer post hoc analyses showed that all three conditions differed significantly from one another (mean difference = 25.9, critical difference = 16.3 for Test-Only vs. Test-With-Feedback; mean difference = 56.2, critical difference = 16.3 for Test-Only vs. Read-Only; mean difference = 30.4, critical difference = 16.3 for Test-With-Feedback vs. Read-Only).

JOL magnitude was not significantly influenced by study condition, $F(2, 89) = .67$, $p = .51$. Final recall accuracy was significantly affected by study condition, $F(2, 89) = 13.97$, $p < .0001$, $\eta_p^2 = .24$. Again, low accuracy in the Test-Only condition largely drove this effect. As in Experiment 1, compared with a test with no feedback, providing feedback had a much larger effect on recall (51.8 vs. 81.2) than it did on JOLs (58.5 vs. 62.5, a difference that was not significant), suggesting that participants may have based their JOLs primarily (or exclusively) on prefeedback cues. (It was not possible to analyze effects of initial test accuracy because there was no initial test.)

Again, we compared recall accuracy in the Read-Only condition with that in the Test-With-Feedback condition. A one-tailed

Table 4
Mean Correlations, Judgments of Learning (JOLs), and Percentage Correct on the Final Test in Experiment 2

	Gamma	JOL	Accuracy
Test only	.805 (.195)	58.5 (19.1)	56.8 (25.3)
Test with feedback	.547 (.271)	62.5 (23.3)	81.2 (15.3)
Read only	.243 (.278)	56.5 (18.8)	78.9 (17.3)

Note. *SD*s are in parentheses.

planned comparison showed no significant testing effect $t(62) = .55$, $p = .29$, a point to which we return in the General Discussion.

In summary, the results replicated Experiment 1. The addition of feedback significantly reduced the gamma correlation between JOLs and recall accuracy, but that correlation remained significantly greater than it was in the Read-Only condition. Comparing the Test-Only and Test-With-Feedback conditions, feedback significantly increased recall accuracy but had no significant effect on JOLs, suggesting that when making JOLs, participants may have been influenced by prior test performance far more than by feedback.

Experiment 3

The only difference between Experiment 2 and Experiment 3 was the learning materials. Students do not frequently study related word pairs (e.g., *thunder-noise*) but they do frequently study foreign language vocabulary. Thus, in Experiment 3, participants studied English-Indonesian translations (e.g., *kiss-ciuman*). There is an inherent difference between these kinds of learning materials: Unlike English word pairs, foreign vocabulary requires learning the target word itself as well as the association between the cue and target. Furthermore, as the results will demonstrate, the translations were more difficult to remember than the pairs from Experiments 1 and 2. Thus, it seemed possible that changing the materials could have a meaningful impact on the pattern of results obtained in the first two experiments.

Method

Participants. Ninety participants (59 women, 31 men; median age = 26 years, range = 18–70 years) were paid $1.00 for completing the experiment, which took 15–20 min. Because of random assignment, there were 28 participants in the Test-Only condition, and 31 participants in each of the other two conditions.

Materials and procedure. Except for a change in learning materials, the procedure in Experiment 3 was identical to the procedure in Experiment 2. The materials consisted of 36 English-Indonesian word pairs (e.g., *Market-Pasar*). The English words were all concrete nouns (Appendix). The Indonesian language was selected because it is unfamiliar to most Americans.

A three-step process was used to translate the words into Indonesian. We first entered a list of English words in Google Translate and selected Indonesian as the target language. We then took the resulting Indonesian words and used Google Translate to translate them back into English. We selected items for use in the study if the original English word was the same as the back-translated English word. In addition, we excluded cognates (e.g., *Truck-Truk*) and words that were too long (e.g., *terbakar sinar matahari-sunburn*).[2]

Results

When analyzing gamma correlations, 6 participants were excluded from the analyses (these participants were not excluded from other analyses), two (from the Test-Only and Read-Only conditions) whose gamma correlations were more than 2.5 *SD*s below the average gamma correlation and four for whom gamma correlations could not be computed, either because their test accuracy was perfect or because they gave the same JOLs for every item. Thus, gamma correlations were analyzed for 25, 29, and 30 participants in the Test-Only, Read-Only, and Test-With-Feedback conditions, respectively.

JOL magnitude, recall accuracy, and resolution are presented in Table 5. Gamma correlations were significantly affected by condition, $F(2, 81) = 10.85$, $p < .0001$, $\eta_p^2 = .21$. Tukey-Kramer post hoc tests confirmed that gamma correlations were significantly higher in the Test-Only condition than in either of the other two conditions (compared with the Test-With-Feedback condition, mean difference = 16.5, critical difference = 15.1; compared with the Read-Only condition, mean difference = 29.5, critical difference = 15.2). The gamma correlation in the Test-With-Feedback condition was not significantly higher than it was in the Read-Only condition (mean difference = 13.1, critical difference = 14.5).

Study condition did not significantly affect JOL magnitude, $F(2, 87) = 1.85$, $p = .16$, but did have a marginally significant effect on final test accuracy, $F(2, 87) = 2.66$, $p = .08$, $\eta_p^2 = .21$. As Table 5 shows, final test accuracy was higher in the Read-Only condition than in the Test-With-Feedback condition—the opposite of a testing effect. This divergence may have occurred because studying was self-paced or because participants were not asked to input their test responses. However, this difference should be interpreted cautiously because it was not significant, according to a Tukey post hoc analysis (mean difference = 6.7, critical difference = 15.2), nor was it replicated in Experiment 1, 2, or 4.

Discussion

Testing enhanced metacognitive accuracy in Experiment 3. Like the previous studies, however, metacognitive accuracy was greater when tests were not followed by feedback than when they were. Indeed, there was not a significant difference between gamma correlations in the Test-With-Feedback and Read-Only conditions in Experiment 3. One reason for this lack of difference appears to be a distinct increase in the gamma correlation in the Read-Only condition relative to the previous studies (although all gamma correlations rose), which may have been because of each foreign language word having distinctive qualities that served as a basis for judgment. There was some hint that participants adjusted their JOLs based on feedback—JOLs were higher in the Test-With-Feedback condition than in the Test-Only condition—but the effect was not significant.

Experiment 4

Results from Experiment 1–3 indicated that providing feedback, in the form of the correct answer, reduced JOL accuracy, when compared with JOLs made based on a cue alone. In Experiment 4 we examined feedback's influence on study decisions. When deciding what to study, people tend to prioritize information that they regard as least well-known (e.g., Son & Metcalfe, 2000), leading to a negative correlation between choosing to study an item and that item's memory strength. Because feedback appears to dimin-

[2] This method revealed that, according to Google Translate, many of the English-Indonesian word-pairs used by Kornell and Son (2009) were inaccurate translations. This inaccuracy was probably immaterial from the perspective of the participants in that study.

Table 5
Mean Correlations, Judgments of Learning (JOLs), and Percentage Correct on the Final Test in Experiment 3

	Gamma	JOL	Accuracy
Test only	.870 (.144)	36.2 (21.2)	34.8 (26.9)
Test with feedback	.705 (.201)	46.2 (19.0)	43.2 (21.3)
Read only	.574 (.311)	42.1 (19.9)	49.8 (26.7)

Note. SDs are in parentheses.

ish metacognitive accuracy, we anticipated that it would diminish the optimality of study decisions as well.

Method

Participants. Eighty-eight participants (59 women, 29 men; median age = 27 years, range = 18–81 years) were paid $1.00 for completing the experiment, which took 15–20 min. They were recruited via Amazon's Mechanical Turk. All participants reported living in the United States.

Procedure. Experiment 4 was identical to Experiment 3 with two exceptions: During the second phase of the experiment, participants were not asked to make a JOL at the end of each trial. Instead, they were asked to select one of two buttons labeled "Study Again" and "Do Not Study Again." They were instructed to select each button roughly half of the time (although many did not honor this request), and were told that they would be allowed to restudy the items they selected before the test. In reality there was no restudy phase (see Finn, 2008; Rhodes & Castel, 2009, for similar procedures). The second difference was that instead of playing Tetris, participants played Asteroids during the distractor task.

Results

Study condition significantly affected the proportion of items participants chose to drop from further studying, $F(2, 85) = 3.29$, $p < .05$, $\eta_p^2 = .07$. As Table 6 shows, participants dropped more items in the Read-Only condition than the other two conditions. If participants had been allowed to restudy, this condition would likely have produced relatively low levels of final recall performance.

Study condition also affected final test accuracy, $F(2, 85) = 3.78$, $p < .05$, $\eta_p^2 = .08$. There was a testing effect—the Test-With-Feedback condition produced the best performance—although based on a post hoc Tukey's test, the only significant difference among the three conditions was between the Test-With-Feedback condition and the Test-Only condition (mean difference = 15.8, critical difference = 13.8).

Table 6
Mean Phi and Gamma Correlations, Mean Number of Items Participants Chose to Stop Studying, and Mean Percentage Correct on the Final Test in Experiment 4

	Phi	Gamma	Chose drop	Accuracy
Test only	.426 (.323)	.703 (.309)	47.8 (22.9)	32.5 (18.0)
Tes with feedback	.204 (.325)	.419 (.548)	47.5 (28.8)	48.3 (25.9)
Read only	.246 (.327)	.597 (.540)	63.3 (28.2)	41.5 (21.1)

Note. *SD*s are in parentheses.

The data from Experiment 4 fit into a 2 × 2 table: Participants requested restudy or not and they answered correctly or not. Thus, these data were analyzed by computing phi correlations for each participant (gamma correlations are also listed for comparison with the other experiments). Twelve participants were excluded for whom correlations could not be computed. The net result was that data from 76 participants were analyzed, with 27, 24, and 25 participants in the Test-Only, Test-With-Feedback, and Read-Only conditions, respectively.

Study condition significantly affected phi correlations, $F(2, 73) = 3.42$, $p < .05$, $\eta_p^2 = .09$. The highest correlation occurred in the Test-Only condition; surprisingly, the Test-With-Feedback condition produced the lowest correlation. Post hoc Tukey-Kramer tests showed that the average phi correlation in the Test-With-Feedback condition was significantly lower than in the Test-Only condition (mean difference = 22.2, critical difference = 21.9), but the Read-Only condition did not differ from the other two conditions.

Choosing to restudy the least well-known items is a common strategy. Another strategy, however, is to choose to drop the most difficult items and restudy the easier items. This behavior typically occurs when people have modest goals or find a task very difficult (Ariel, Dunlosky, & Bailey, 2009; Kornell & Metcalfe, 2006; Thiede & Dunlosky, 1999). Restudying difficult items leads to a positive correlation between the choice to drop an item and subsequent test performance; restudying easier items leads to a negative correlation. Twelve of 76 participants had negative Phi correlations (the same participants also had negative gamma correlations) in Experiment 4, suggesting that at least some of these participants may have chosen the "drop difficult items" strategy. When participants with negative correlations were removed from the analysis, however, the pattern of phi correlations was the same: The Test-With-Feedback condition produced the lowest correlations, and the Test-Only condition produced the highest correlations.

Table 7 displays the percentage of items that fell into each cell of the 2x2 data table. These values were computed with data for all participants pooled. Based on the hypothesis that participants mostly ignore feedback, one would predict that in the Test-With-Feedback condition, they should be prone to a certain kind of error: They should choose to restudy items that they ended up answering

Table 7
Percentage of Items Falling Into Each of Four Possible Response Combinations, Based on Study Decision and Final Test Accuracy, Pooled Across Participants in Experiment 4

	Drop	Restudy
Test only		
Incorrect	23	45
Correct	25	7
Test with feedback		
Incorrect	20	31
Correct	27	21
Read only		
Incorrect	32	27
Correct	31	10

correctly (because the feedback meant they did not really need to restudy those items). Such errors happened much more frequently in the Test-With-Feedback condition (21% of trials) than in the other two conditions. In the Test-Only condition, by contrast, participants were highly likely (45% of trials) to choose to restudy items that they had not mastered (i.e., that were not correctly recalled). These findings suggest, consistent with the prior experiments, that in the Test-With-Feedback condition, participants failed to account for the learning that occurred as a result of feedback and based their study decisions on the self-tests they took prior to feedback.

Discussion

The results of Experiment 4 suggest that study decisions were more accurate in the Test-Only condition than the Test-With-Feedback condition. Indeed, restudy choices in the Test-With-Feedback condition were no more accurate than those in the Read-Only condition. These data are consistent with the hypothesis that participants largely ignored feedback when making study decisions, instead basing their decisions on prior test performance.

General Discussion

A large body of research shows that tests enhance metacognitive accuracy (cf. Rhodes & Tauber, 2011a). The present study is the first in which participants took a test, saw the correct answer, and then made a JOL. This procedure is realistic because when people test themselves they customarily check the correct answer afterward, and make ongoing judgments of learning, at least implicitly, as they do so. The results suggest that previous studies may have overestimated the metacognitive benefit of tests.

Tests without feedback produced very high metacognitive accuracy (averaged across Experiment 1–3, $\gamma = .81$). Adding feedback following the tests reduced metacognitive accuracy ($\gamma = .57$), although not to the level of Read-Only trials ($\gamma = .39$). Test-Only trials also produced the highest metacognitive accuracy when participants made study decisions (as measured using Phi correlations in Experiment 4); Test-With-Feedback and Read-Only trials did not differ significantly.

There were two unusual aspects of the procedures: In correspondence with how students usually study, all trials were self-paced and, except in Experiment 1, participants did not make overt responses as they studied. The fact that JOLs were more accurate in the Test-Only condition than the Read-Only condition in all experiments, replicating the dJOL effect, suggests that basic judgment processes were not adversely affected by these procedural deviations.

There was also an unusual aspect of the sample: Participants in the first three studies ranged in age from 18–74. Yet in a meta-analysis of 112 gamma correlation effect sizes, Rhodes and Tauber (2011a) found that 93 came from college students and only eight came from older adults (the others came from children). Thus, most dJOL studies have included primarily college students. We analyzed the potential effects of age on judgments, combining data from Experiments 1–3 to maximize statistical power. A total of 327 participants were included (one was excluded because he did not report his age). The mean age was 32.0 years. Participants were split into older ($n = 148$, mean age = 42.5) and younger ($n = 179$, mean age = 23.3) age groups based on the median of 28 years. We conducted analyses of variance (ANOVAS) with age group and condition as independent variables, using mean JOL, recall, and gamma correlation as dependent variables. Age group had no significant main effects, or interactions with condition, for any variable. Thus, it appears that the findings from these studies can be generalized across age groups.

Testing has been endorsed as a practical way to enhance metacognitive monitoring. The present findings suggest that in realistic situations, when tests are followed by feedback, tests *do* enhance metacognitive monitoring, compared with Read-Only trials, but not as much as previously thought. We found no evidence that they enhance study decisions. Such findings have important theoretical implications, to which we turn next.

Determinants of JOLs

A test-only JOL is made on the basis of an attempt to retrieve information from long-term memory. Feedback makes additional information available, but prior evidence suggests that people tend to underweight, or even ignore, feedback when making judgments. Instead, people base metacognitive judgments on their memory for a past test (see Finn & Metcalfe, 2008). The MPT heuristic is consistent with Koriat's (1997) cue-utilization model, which states that mnemonic cues, such as test performance, are weighted heavily in metacognitive judgments.

According to the MPT heuristic, feedback should have little or no influence on mean JOL levels. Averaged across Experiments 1–3, the mean JOL was 53.2 in the Test-Only condition and 57.0 in the Test-With-Feedback condition. This nonsignificant difference of 3.8 percentage points suggests that participants were only slightly influenced by feedback, if at all. By contrast, the difference in recall accuracy, averaged across Experiments 1–3, was roughly 20 percentage points (50.9 vs. 71.0).[3] This difference was even more pronounced when examining only incorrect answers (see Experiment 1).

In short, the evidence suggests that when making JOLs, participants may have showed a slight sensitivity to feedback, but it was dwarfed by the effect of feedback on learning. Feedback appeared to create an "illusion of not knowing," as knowledge increased but judgments did not. This illusion helps explain why the correlation between JOLs and recall accuracy was lower in the Test-With-Feedback condition than the Test-Only condition. These data also echo the finding that people exhibit a "stability bias" about their own memories, failing to take future learning into account (e.g., Kornell & Bjork, 2009).

The failure to take feedback into account can also explain the results of Experiment 4. Participants who underappreciate the value of feedback should choose to restudy an item that they answered incorrectly during the study phase, even if, without restudying, they would be able to answer it correctly on the final test. That is, participants who were provided with feedback should have chosen to restudy many items that they eventually answered

[3] It is worth noting that the average JOL levels were not insensitive to the likelihood of recall, as demonstrated by the positive gamma correlations. In addition, an ANOVA comparing Experiments 1–3 showed that JOLs were significantly lower in Experiment 3, which used relatively difficult materials, than in the other two experiments.

correctly. Indeed, 21% of items chosen for restudy in the Test-With-Feedback condition were recalled correctly (see Table 7). Such items should have occurred less often when there was no feedback to undervalue, and they did; they accounted for 7% and 10% of trials in the Test-Only and Test-With-Feedback conditions, respectively. Consistent with Experiment 1, it seems likely that the participants particularly underappreciated the value of feedback following incorrect responses.

Collectively, the findings suggest that an underappreciation of the value of feedback decreased the accuracy of metacognitive judgments and study choices. Participants appeared to rely heavily on the MPT heuristic, although evidence from Experiment 1 suggested that they were also sensitive to other, as yet unspecified, cues.

Testing Effects

Participants in the Test-With-Feedback condition did not consistently outperform participants in the Read-Only condition. There was a marginally significant benefit of testing in Experiment 1 and no significant advantage in the other three experiments. One possible explanation of this relatively weak testing effect is the fact that the tests took place shortly after the study phase. A number of studies have shown that testing effects are most pronounced after a longer retention interval (e.g., Roediger & Karpicke, 2006a). This explanation seems questionable, however, because the comparison of interest is between the Read-Only and Test-With-Feedback conditions. Recent research suggests that when tests are followed by feedback, their advantage (or lack thereof), when compared with read-only trials, remains fairly constant over time (Kornell et al., 2011).

A more likely explanation is that participants made JOLs while they studied. Recent evidence suggests testing effects can diminish or disappear if the learning task includes making JOLs (Jönsson, Hedner, & Olsson, 2012). Furthermore, making a JOL can act like a test trial in the way it enhances memory (Sundqvist, Todorov, Kubik, & Jönsson, 2012). The present findings did not directly address this issue by including a no-JOL condition, but they are consistent with the idea that making JOLs enhances memory for presentation trials more than it does for test trials, thereby diminishing the testing effect.

Memory or Metamemory?

Kimball and Metcalfe (2003) eliminated the dJOL effect by providing feedback following JOLs. In contrast, in the present studies, the dJOL effect was reduced but remained significant. The reason Kimball and Metcalfe found a larger reduction in JOL accuracy as a result of feedback may be artifactual. On some trials, their participants might have wanted to adjust their JOLs upon receiving feedback, either upward (after being surprised to find out that they had given a correct answer) or downward (after being surprised to find their answer had been wrong). They were not given the opportunity to do so. Under normal circumstances, learners are free to make such adjustments as they study.

Like Kimball and Metcalfe (2003), we endeavored to test the memory hypothesis as an explanation of the dJOL effect. As noted previously, the memory hypothesis posits that tests increase gamma correlations because recalled items, which are associated with high JOLs, become more memorable as a result of the test, not because tests produce accurate JOLs. The memory hypothesis hinges on the idea that some tested items benefit from tests but others do not. Providing feedback entails that all items benefit from being tested, even those that are not recalled (see Kornell et al., 2011). This eliminates the differential treatment of tested versus nontested items. According to the memory hypothesis, therefore, the relative accuracy of JOLs should be roughly equal in the Read-Only and Test-With-Feedback conditions. Our data consistently showed that it was not equal. Thus, the current results cannot be fully explained by the memory hypothesis; at least some of the enhancement in JOL accuracy because of tests with feedback can be attributed to metamemory, a finding at variance with Kimball and Metcalfe's (2003) conclusion. Indeed, our data are at least partially consistent with the monitoring-dual-memories hypothesis (e.g., Dunlosky & Nelson, 1992), according to which metacognitive accuracy is enhanced when judgments can be made based on a meaningful attempt to recall information from long-term memory.

Whether or not memory effects played a role in JOL accuracy cannot be determined. For example, memory differences may explain why the correlations in the Test-Only condition exceeded the correlations in the Test-With-Feedback condition. However, the fact that participants seemed to ignore the value of feedback may be sufficient to explain this difference. The bottom line is that metamemory effects did play a role and memory effects might have contributed as well.

Conclusion

Overall, the experiments we reported suggest a hidden danger of feedback following a test: it makes metacognitive judgments less accurate, although not as inaccurate as judgments based on studying without any test. One outcome of this danger is that learners might shift study time away from items that were answered correctly on a prior test. People tend to assume that once an item has been recalled it will not be forgotten (Koriat, Bjork, Sheffer, & Bar, 2004) and stop studying items they have answered correctly a single time (Kornell & Bjork, 2008). When making study decisions, it is natural to focus on items that one has struggled with, but it is important to realize that memory is anything but stable (cf. Kornell, 2012): items that were answered correctly can be forgotten, and items that were not answered correctly at one point may have been learned during subsequent studying.

What can students learn from these studies? First, the optimal way to study, at least for information that lends itself to memorization, may be to test oneself and then obtain feedback. Doing so produced the most learning in the present studies. But there is a hidden cost of feedback: Adding feedback after a test decreases metacognitive accuracy. Second, most learners base judgments on how they did on a prior test, even after they receive feedback, but feedback following an error has a large effect on learning that should not be ignored. Thus, tests with feedback enhance learning, but students should be aware of the surprising power of feedback for learning and its potential adverse effects on monitoring.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgments-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39,* 171–184. doi:10.3758/s13421-010-0002-y

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138,* 432–447. doi:10.1037/a0015928

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41,* 28–35. doi:10.3758/s13421-012-0249-6

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64,* 467–484. doi:10.1080/17470218.2010.502239

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed JOL effect. *Memory & Cognition, 20,* 374–380. doi:10.3758/BF03210921

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565. doi:10.1006/jmla.1994.1026

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36,* 813–821. doi:10.3758/MC.36.4.813

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 238–244. doi:10.1037/0278-7393.33.1.238

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58,* 19–34. doi:10.1016/j.jml.2007.03.006

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10,* 597–602. doi:10.3758/BF03202442

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40,* 505–513. doi:10.3758/s13421-011-0174-0

Hartwig, M. K., & Dunlosky, J. D. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19,* 126–134. doi:10.3758/s13423-011-0181-y

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17,* 797–801. doi:10.3758/PBR.17.6.797

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology, 59,* 251–257. doi:10.1027/1618-3169/a000150

Karpicke, J. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486. doi:10.1037/a0017341

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31,* 918–929. doi:10.3758/BF03196445

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93,* 329–343. doi:10.2307/1422236

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual CHI conference on human factors in computing systems–CHI ' 08,* 453. doi:10.1145/1357054.1357127

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. doi:10.1037/0096-3445.126.4.349

Koriat, A. (1998). Illusions of knowing: The link between knowledge and metaknowledge. In V. Y. Yzerbyt & G. Lories (Eds.), *Metacognition cognitive and social dimensions* (pp. 16–34). London, United Kingdom: Sage.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643–656. doi:10.1037/0096-3445.133.4.643

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. doi:10.1037/0096-3445.135.1.36

Kornell, N. (2012). A stability bias in human memory. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 4–7). Boston, MA: Springer. doi:10.1007/978-1-4419-1428-6

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. doi:10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits-and costs-of dropping flashcards. *Memory, 16,* 125–136. doi:10.1080/09658210701763899

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138,* 449–468. doi:10.1037/a0017350

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. doi:10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 989–998. doi:10.1037/a0015729

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 609–622. doi:10.1037/0278-7393.32.3.609

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501. doi:10.1080/09658210902832915

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15,* 307–318. doi:10.1037/a0017599

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 671–685. doi:10.1037/a0018785

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23. doi:10.3758/s13428-011-0124-6

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Retrieved from http://w3.usf.edu/FreeAssociation/

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. doi:10.1037/0033-2909.95.1.109

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116. doi:10.1037/0003-066X.51.2.102

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2,* 267–270. doi:10.1111/j.1467-9280.1991.tb00147.x

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.125–173). New York, NY: Academic Press. doi:10.1016/S0079-7421(08)60053-5

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.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8. doi:10.1037/0278-7393.31.1.3

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19,* 559–579. doi:10.1080/09541440701326022

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16,* 550–554. doi:10.3758/PBR.16.3.550

Rhodes, M. G., & Tauber, S. K. (2011a). The influence of delaying Judgments of Learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137,* 131–148. doi:10.1037/a0021705

Rhodes, M. G., & Tauber, S. K. (2011b). Eliminating the delayed JOL effect: The influence of the veracity of retrieved information on metacognitive accuracy. *Memory, 19,* 853–870. doi:10.1080/09658211.2011.613841

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* 243–257. doi:10.1037/a0016496

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

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. doi:10.1037/0278-7393.26.1.204

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3,* 315–316. doi:10.1111/j.1467-9280.1992.tb00680.x

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 95–114). New York, NY: Psychology Press.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43,* 155–167. doi:10.3758/s13428-010-0039-7

Sundqvist, M. L., Todorov, I., Kubik, V., & Jönsson, F. U. (2012). Study for now, but judge for later: Delayed judgments of learning promote long-term retention. *Scandinavian Journal of Psychology, 53,* 450–454. doi:10.1111/j.1467-9450.2012.00968.x

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' Judgments-of-learning (JOLs) in multitrial learning. *Psychology and Aging, 27,* 474–483. doi:10.1037/a0025246

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. doi:10.1037/0278-7393.25.4.1024

Weaver III, C. A., & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1058–1065. doi:10.1037/0278-7393.29.6.1058

Appendix

Word Pairs Used as Stimuli in Experiments 3 and 4

Cue	Target	Cue	Target
Army	Tentara	Jelly	Jeli
Baby	Bayi	King	Raja
Blood	Darah	Kiss	Ciuman
Body	Tubuh	Lake	Danau
Bouquet	Buket	Magazine	Majalah
Breast	Payudara	Market	Pasar
Cane	Tebu	Monk	Biarawan
Church	Gereja	Nail	Kuku
Cigar	Cerutu	Paper	Kertas
Corn	Jagung	Ship	Kapal
Corpse	Mayat	Singer	Penyanyi
Cradle	Buaian	Sugar	Gula
Daffodil	Bakung	Swamp	Rawa
Diamond	Berlian	Temple	Candi
Fisherman	Nelayan	Tweezers	Pinset
Gold	Emas	Umbrella	Payung
Hammer	Palu	Vehicle	Kendaraan
Horse	Kuda	Village	Desa

Received June 1, 2012
Revision received September 24, 2012
Accepted December 21, 2012 ■