Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 1, 238–244

Copyright 2007 by the American Psychological Association
0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.1.238

The Role of Memory for Past Test in the Underconfidence With Practice Effect

Bridgid Finn and Janet Metcalfe
Columbia University

According to the Memory for Past Test (MPT) heuristic, judgments of learning (JOLs) may be based, in part, on memory for the correctness of answers on a previous test. The authors explored MPT as the source of the underconfidence with practice effect (UWP; A. Koriat, L. Sheffer, & H. Ma'ayan, 2002), whereby Trial 1 overconfidence switches to underconfidence by Trial 2. Immediate and delayed JOLs were contrasted because only immediate JOLs demonstrate UWP. Consistent with MPT for immediate JOLs, Trial 1 test performance better predicted Trial 2 JOLs than did Trial 2 test performance. Delayed JOLs showed the reverse. Furthermore, items forgotten on Trial 1 but remembered on Trial 2 contributed disproportionately to UWP, but only with immediate JOLs.

Keywords: metacognition, underconfidence, judgments of learning, memory for past test

Research has shown that judgments about upcoming test performance generally exhibit overconfidence, whereby they are higher on average than mean test performance (Ayton & McClelland, 1997). Whereas this finding is widely but not always found (Fischhoff, Slovic, & Lichtenstein, 1977), the conclusion that people are usually overconfident is at odds with the findings of Koriat (1997) and Koriat, Sheffer, and Ma'ayan (2002) showing that overconfidence is confined to the first trial in a multitrial learning situation. Judgment inaccuracy, seen in the first trial as overconfidence, shifts toward underconfidence in Trial 2 and beyond. This phenomenon has been dubbed the Underconfidence with Practice (UWP) effect.

Although the UWP effect appears to be robust—with one exception that will be detailed shortly—the causes of the effect are currently unknown. Koriat et al. (2002) demonstrated that it persisted despite a variety of experimental manipulations. For example, feedback about performance on a prior trial failed to diminish underconfidence on the next trial. Both incorrectly and correctly recalled Trial 1 items showed underconfidence on Trial 2. The UWP effect obtained with both fixed and self-paced study-time allocation. It does not appear to be attributable simply to subjective control of study. With incentives given for accurate judgments, a manipulation that increased Trial 1 judgment of learning (JOL) accuracy, UWP persisted. The UWP effect arose in both item-by-item and aggregate JOLs. Related and unrelated word pairs both showed Trial 2 underconfidence, suggesting that UWP is not materials dependent. Item difficulty has also been varied. Numerous studies have shown that easy materials tend to result in less overconfidence than difficult materials (Lichtenstein & Fischhoff, 1977). However, Koriat et al. (2002) reported that easy and difficult items showed UWP, casting some doubt on the idea that the UWP effect reflects a hard–easy difference whereby as items become easier from trial to trial (with learning), underconfidence increases.

Finally, Serra and Dunlosky (2005) recently provided data against the idea that UWP is attributable to retrieval fluency. The idea is that although people are likely to recall on Trial 2 items that they already recalled on Trial 1, they may feel less than 100% confident about their ability to do so, especially for items that were recalled with difficulty on the first trial. Thus, the items that lacked fluency on the first trial recall—even though recall was eventually successful—should be treated gingerly and given relatively low JOLs. If the retrieval fluency explanation were correct, those items should contribute disproportionately to the UWP effect. Despite the reasonableness of the hypothesis, Serra and Dunlosky's data provided no support for this potential locus of the UWP effect.

The one procedure, so far, that has modulated the UWP effect is delaying the JOLs (Koriat & Ma'ayan, 2005; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Meeter & Nelson, 2003; Scheck & Nelson, 2005; Serra & Dunlosky, 2005). The original UWP proposal was based on studies that all used immediate JOLs. Studies that have contrasted immediate with delayed judgments have consistently found that immediate Trial 2 JOLs show more underconfidence than delayed JOLs. Indeed, the data to date have suggested that there may be no UWP effect with delayed JOLs. Meeter and Nelson (2003) showed only a 1% difference between delayed judgments and recall performance. Serra and Dunlosky (2005) showed underconfidence for both delayed and immediate judgments but a greater degree of underconfidence in the immediate condition. Data reported from earlier work from the same lab (Dunlosky & Connor, 1997) showed slight overconfidence with delayed JOLs. Koriat and Ma'ayan (2005) and Koriat et al. (2006) also reported slight overconfidence on Trial 2 delayed judgments. Taken together these reports suggested that delayed JOLs are not underconfident.

Bridgid Finn and Janet Metcalfe, Department of Psychology, Columbia University.

This research was supported by National Institute of Mental Health Grant RO1-MH60637. We thank Niall Bolger, Matt Greene, Rainer Romero, and Nate Kornell for their help and comments.

Correspondence concerning the article should be addressed to Janet Metcalfe, Department of Psychology, Columbia University, 401B Schermerhorn Hall, New York, NY 10027. E-mail: jm348@columbia.edu

One focus of research likely to be fruitful is the exploration of the information people use to make immediate JOLs as compared with delayed JOLs because the former show UWP and the latter do not. Presumably, the heuristic information that is used to make immediate as compared with delayed JOLs may reveal the basis of the UWP effect. We adopted this strategy here.

In reviewing the literature contrasting immediate and delayed JOLs, it seems clear that people use different heuristics when they make these two kinds of JOLs (but see Sikström & Jönsson, 2005). It is well established that delayed JOLs show a predictive accuracy advantage over immediate JOLs, an advantage that may modulate metacognitive biases. There is disagreement about whether the improved accuracy with delayed JOLs is a memory effect due to retrieval (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992) or a pure metamemory effect (Koriat, 1997; Nelson & Dunlosky, 1991; Nelson, Dunlosky, Graf, & Narens, 1994), but there is general agreement on the idea that the information people use to make delayed JOLs involves a target retrieval attempt that is highly diagnostic of their subsequent target retrieval attempt at time of test.

There is less agreement about what information people use to make immediate JOLs. Information in short-term memory (Nelson & Dunlosky, 1991), an ease of processing evaluation that takes into account ease of encoding (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat & Ma'ayan, 2005), or normative difficulty of the items (Koriat, 1997) have been proposed as candidates. This information does not engender accurate JOLs, however, as is reflected by very low first trial gamma correlations. This information may be used primarily in the first trial, whereas other heuristics may come into play by the second trial (see Koriat et al., 2006).

We propose that after a test people may make subsequent immediate JOLs on the basis of the outcomes of the test. If they remember getting the item right on the test, they give it a high JOL; if they remember that they did not get it right—despite the presence of that item in their current working memory and despite the learning that may have occurred—they may tend to give it a low JOL. We call this heuristic Memory for Past Test (MPT).

Our proposal that people may use the MPT heuristic (when such memory is available—i.e., by the second trial—in the immediate but not the delayed JOL condition) contrasts with Koriat et al. (2002), who suggested that previous test trial performance is unimportant for the UWP effect. They argued against the possible effect of previous-trial test performance because recalled and unrecalled items both showed the UWP effect. Underconfidence for the unrecalled items follows straightforwardly from the MPT heuristic. Underconfidence for the recalled items would be expected if there were any forgetting at all of what one recalled, or indeed, if there were any variability in those judgments, because the judgments are constrained to not go above 100%. Variability cannot be normally distributed, however, and all variability will be downward from 100% and, necessarily, will register as underconfidence. Despite Koriat et al.'s (2002) argument, it seems that a reliance on past test performance might be central to the UWP effect.

In the face of nondiagnostic information during the relatively inaccurate immediate JOL assessment, prior test performance on individual items may serve as one of the few reliable indicators of upcoming test performance and might well be used. Insofar as other, more reliable information is available—such as the goodness of a current retrieval attempt—as is available with delayed JOLs, it seemed plausible to suppose that judgment makers may be less dependent on prior test information. Accordingly, it is plausible that JOLs will rely on MPT when the judgments are made immediately but not at a delay.

An underlying assumption, for the MPT hypothesis, is that people can remember fairly well how they performed on individual items on their past test. Research has demonstrated that people have very good memory for their previous test performance (Gardiner & Klee, 1976), and memory performance on the prior trial is strongly correlated with JOL ratings on a subsequent trial (Hertzog, Dixon, & Hultsch, 1990; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984; Thiede, 1999). Furthermore, several experiments have shown that a prior test can result in modification of subsequent-trial encoding strategies (Gardiner, Passmore, Herriot, & Klee, 1977; Halff, 1977; LaPorte & Voss, 1974). So, people appear to have access to their previous-trial test performance and to be able to use this information.

Basing one's current judgments, in part or in whole, on previous test outcomes is an empirical possibility and one that could lead to systematic bias in the direction of underconfidence. UWP would result if such a reliance on the item-by-item outcome of the previous test was either the exclusive heuristic used—because, if any learning occurs over trials, Trial 1 performance will be lower than Trial 2 performance—or if it were coupled with any discounting of or inaccuracy in the amount of learning due to current-trial study efforts.

Experiment 1

Our hypothesis was that the UWP effect, which should occur selectively in the immediate JOL condition but not the delayed JOL condition, is due in part to a reliance on the MPT heuristic. When judgments can rely on more diagnostic prospective information, as with delayed JOLs, MPT need not contribute to the judgment. We hypothesized that Trial 2 immediate JOLs would be better predicted by Trial 1 test than would Trial 2 delayed JOLs because of the proposed selective use of the MPT heuristic in the immediate case. We also hypothesized that items that were forgotten on Test 1 but remembered on Test 2 should contribute disproportionately to the UWP effect in the immediate JOL condition as compared with the delayed JOL condition, again, because of reliance on the MPT heuristic in the immediate JOL condition.

Method

Participants. The participants were 24 undergraduates, ages 18 to 24 years, from Columbia University and Barnard College. They received partial course credit for their participation. Participants were treated in accordance with APA ethical guidelines.

Design. The experiment consisted of a 2 (trial: 1 or 2) × 2 (JOL condition: delayed or immediate) × 2 (order of immediate JOL: first or second) within-participants, between-list design, with 48 word pairs per JOL condition.

Materials and apparatus. The lists consisted of 48 cue–target pairs randomly drawn without replacement, from the Toronto Word Pool, a pool of 1,080 common English two-syllable words

(Friendly, Franklin, Hoffman, & Rubin, 1982). Mean length of the cue and target was 6.24 letters, and no words exceeded 8 letters. Selection of which list would be slated for immediate or delayed JOLs was determined in a randomized counterbalanced order for participants.

Procedure. Participants were instructed that they would be studying 48 cue–target pairs, making JOLs, taking a cued-recall test, and that they would go through this procedure twice. JOLs were explained as judgments of learning based on what participants thought their chances were for recalling the second word when given the first word during a memory test that would happen in a few minutes, and they were asked to use a scale from 0%–100%. In the immediate condition, participants were informed that they would make a JOL immediately after they studied the pair and after the entire study list had been presented in the delayed condition. JOLs were made with only the cue present. After completing two study-test trials on the same list of words (presented in a different random order), participants learned a new list and were given instructions about how to make the other judgment (either immediate or delayed, depending on what they had been presented for the first list). Order was counterbalanced but had no effect in any analysis and will not be further discussed.

Cue–target pairs were shown for 3 s each. In the immediate condition, pairs were immediately followed by presentation of the cue for an immediate JOL. In the delayed condition, each pair was followed by a new cue–target study presentation. After all of the pairs had been presented, approximately 2.5 min later, the cues were reshuffled and presented for delayed judgments.

A self-paced cued-recall test followed approximately 10 min later. The second trial followed approximately 3 min after the Trial 1 test. Participants completed two trials for each JOL condition.

Results

Trial 1 recall performance means were .22 for immediate and .11 in the delayed condition; for Trial 2, the means were .40 and .31, respectively. Recall performance increased from Trial 1 to Trial 2, $F(1, 23) = 85.37$, $MSE = .01$, $p < .001$, $\eta^2 = .78$ (effect size was computed with partial eta squared), and there was a difference between the conditions, $F(1, 23) = 12.41$, $MSE = .02$, $p = .002$, $\eta^2 = .35$. Immediate JOL items had higher recall performance than did delayed JOL items. There was no interaction between trials and condition ($F < 1$).

Trial 1 JOL means were .37 for immediate and .20 for delayed; Trial 2 means were .35 and .32, respectively. JOLs slightly increased over trials; mean Trial 1 JOL was .28 and mean Trial 2 JOL was .33, but this trend was not significant. There was a main effect of condition, $F(1, 23) = 10.67$, $MSE = .02$, $p = .003$, $\eta^2 = .32$; immediate JOLs were higher overall. The interaction for JOLs between trials and condition was significant, $F(1, 23) = 21.33$, $MSE = .01$, $p < .001$, $\eta^2 = .48$.

The UWP effect is a measure of absolute bias, corresponding to the overall mean of the judgments as compared with the mean recall score. This can be measured in terms of calibration, which is the mean recall performance subtracted from the mean JOL. An underconfidence bias is evidenced if the mean JOL underestimates mean recall performance. In contrast, the relative accuracy, or resolution, of JOLs is a different metacognitive assessment of whether the person can predict which items will be remembered and which will be forgotten. Resolution differs from calibration because it measures whether the relative magnitude of JOL values is appropriate: Do people get highly ranked JOLs correct on a test? When JOLs are higher for items that will be answered correctly and lower for items that will be answered incorrectly, resolution is high.

The UWP effect is a calibration effect. However, the MPT hypothesis indicates that the UWP effect may be a result of the reliance on individual items' past recall. This is measured by the relative accuracy or resolution not to the upcoming test but rather to the past test. According to the MPT heuristic, people assign items that were correctly recalled on the last test high JOLs and items that were forgotten low JOLs, resulting in high resolution between current JOLs and past (but not necessarily upcoming) test. Thus, we are concerned with both calibration (the indicator of the UWP effect) and resolution (especially with respect to past test—an indicator of the use of the MPT heuristic) in the analyses that follow.

We first report calibration. A main effect of trial, $F(1, 23) = 52.36$, $MSE = .01$, $p < .001$, $\eta^2 = .70$, showed overconfidence exhibited on Trial 1 ($M = .12$), $t(23) = 4.14$, $p < .001$, as measured by a single-sample t test; on Trial 2, calibration was not significantly different from zero ($M = –.02$; $t < 1$). On Trial 1, both immediate ($M = .15$), $t(23) = 3.89$, $p = .001$, and delayed ($M = .09$), $t(23) = 2.80$, $p = .01$, JOLs demonstrated significant overconfidence. On Trial 2, immediate JOLs were significantly underconfident, $t(23) = 2.81$, $p = .03$; the mean calibration score was –.06. Delayed JOL's calibration score on Trial 2 was .01 and was not significantly different from zero ($t < 1$). Of interest was the large and significant Trial × Condition interaction, $F(1, 23) = 10.51$, $MSE = .01$, $p = .004$, $\eta^2 = .31$. Post hoc tests showed a large difference between the two conditions on Trial 2, $t(23) = 2.78$, $p = .01$.

Next, we report resolution: If MPT was being used then items that were recalled on Trial 1 test should have high Trial 2 JOLs, and items that were not recalled on Trial 1 test should have low Trial 2 JOLs. If MPT was not being used, and Trial 2 JOLs were really based on the improved Trial 2 performance, then correct Trial 2 items should show high Trial 2 JOLs, whereas incorrect Trial 2 items should show low Trial 2 JOLs. These two correlations—between Trial 1 test and Trial 2 JOLs and between Trial 2 test and Trial 2 JOLs—can be computed for each participant. Because the tests are themselves correlated, though, simultaneous multiple regression is indicated. See Table 1 for the Pearsons and Gamma intercorrelations between each measure for the immediate and delayed JOL conditions.

Fitting such a multiple linear regression model to each participant's data in both the immediate and delayed conditions affords a powerful test of our predictions about the factors affecting Trial 2 JOLs. Multiple linear regression is a technique used to evaluate the independent contributions of two or more predictors on a dependent measure (Cohen, Cohen, West, & Aikin, 2003), in this case Trial 2 JOLs. Separate simultaneous

Table 1
Intercorrelations Between All Measures Reported in Experiments 1 and 2 for Immediate and Delayed Conditions

	Immediate				Delayed			
	Pearsons		Gamma		Pearsons		Gamma	
Intercorrelation	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
JOL Trial 1–JOL Trial 2	.27	.14	.24	.14	.41	.35	.51	.50
Recall Trial 1–Recall Trial 2	.49	.35	.92	.88	.46	.37	.97	.91
JOL Trial 1–Recall Trial 1	.21	.19	.33	.33	.69	.75	.80	.89
JOL Trial 2–Recall Trial 2	.49	.25	.64	.36	.71	.71	.81	.85
JOL Trial 1–Recall Trial 2	.19	.05	.28	.09	.37	.31, *ns*	.53	.56, *ns*
JOL Trial 2–Recall Trial 1	.66	.55	.86	.81	.46	.34	.81	.74

Note. All data, unless noted, are significantly different from zero at $p \leq .01$. JOL = judgment of learning.

models[1] were computed for immediate and delayed conditions, yielding four standardized beta coefficients for each participant. Each participant's standardized beta coefficients indicate how well the predictor variables (Trial 1 test or Trial 2 test) accounted for the JOL, that is, the extent to which Test 1 recall success or failure (MPT) or Test 2 recall success or failure was used to make JOLs. The standardized beta coefficients for each participant, which resulted from their own simultaneous multiple regressions, were then entered into a 2 × 2 analysis of variance (ANOVA), with factors being condition (immediate or delayed) and test beta (1 or 2).

As is shown in Figure 1, immediate JOLs showed a larger mean standardized beta coefficient for Trial 1 test than for Trial 2 test, as predicted by the MPT hypothesis, whereas the delayed condition showed the reverse pattern. This interaction was significant, $F(1, 18) = 27.51$, $MSE = .13$, $p < .001$, $\eta^2 = .60$. Post hoc tests demonstrated that Trial 1 test βs and Trial 2 test βs were significantly different in the immediate JOL condition, $t(18) = 2.79$, $p = .01$, and in the delayed JOL condition, $t(18) = 5.45$, $p < .001$. There was no main effect of either test trial or JOL condition.[2]

Finally, we investigated whether unrecalled items on Test 1 that were subsequently remembered on Test 2 contributed disproportionately to the UWP effect. If MPT were being used, as we expect is the case in the immediate but not the delayed condition, then JOLs for items that were forgotten on Trial 1 but remembered on Trial 2 (FR) should be lower than items that were remembered on both trials (RR). Furthermore, the selective underconfidence on the FR items should be amplified in the immediate JOL condition, as compared with the delayed JOL condition.

We computed a 2 (JOL condition: immediate vs. delayed) × 2 (item status: FR vs. RR) within-participants ANOVA. The mean immediate and delayed JOLs for FR and RR are depicted in Figure 2. There was a significant interaction between JOL condition and item status, $F(1, 20) = 7.22$, $MSE = .03$, $p = .01$, $\eta^2 = .27$. Immediate and delayed RR conditions were not significantly different from one another ($t < 1$). The locus of the interaction was in the FR conditions: JOLs were disproportionately low in the immediate JOL conditions with the items that had been forgotten on the previous test. The difference between FR and RR was over twice as large in the immediate condition (.33) as compared with the delayed condition (.14), $t(20) = 2.69$, $p = .01$. The immediate FR condition, which showed the lowest mean JOL overall (.40), was significantly different from the delayed FR condition, $t(21) = 2.67$, $p = .01$.

Discussion

An ANOVA using the standardized beta coefficients from each participant's simultaneous regression model provided evidence that immediate JOLs rely on prior test, as indicated by the MPT heuristic. When judgments were immediate, Trial 1 test best predicted an item's Trial 2 JOL. When judgments were delayed, however, Trial 2 test best predicted Trial 2 JOLs, indicating that people did not rely on the MPT heuristic in this condition. We also showed that in comparison with delayed JOLs, immediate JOLs were selectively underconfident on those particular items that had been forgotten on the last test. These results suggest that people use the MPT heuristic in making Trial 2 immediate JOLs and that the use of this heuristic contributes to the underconfidence seen, selectively, in that condition.

Experiment 2

We tested the reliability of the results in Experiment 1 by performing a second experiment, with some variations, targeting the same questions. In Experiment 2, participants made delayed and immediate JOLs within list rather than between list. Because both conditions were mixed within a single list, this was a more demanding test of whether the sources of information with the two kinds of judgments might be different. Nevertheless, we expected that the results would be similar to those found in Experiment 1.

Method

Participants. Thirty Columbia University and Barnard College undergraduates, ages 18 to 26 years, participated for course credit or cash.

[1] We used a simultaneous regression model (as opposed to hierarchical) because of considerable correlation between Test 1 and Test 2 recall performance, which artificially reduces the contribution of whichever variable is entered second into the hierarchical model, and because we needed to test the relative contributions of Trial 1 test or Trial 2 test.

[2] We found the identical pattern with a nonparametric test using ranked JOLs.

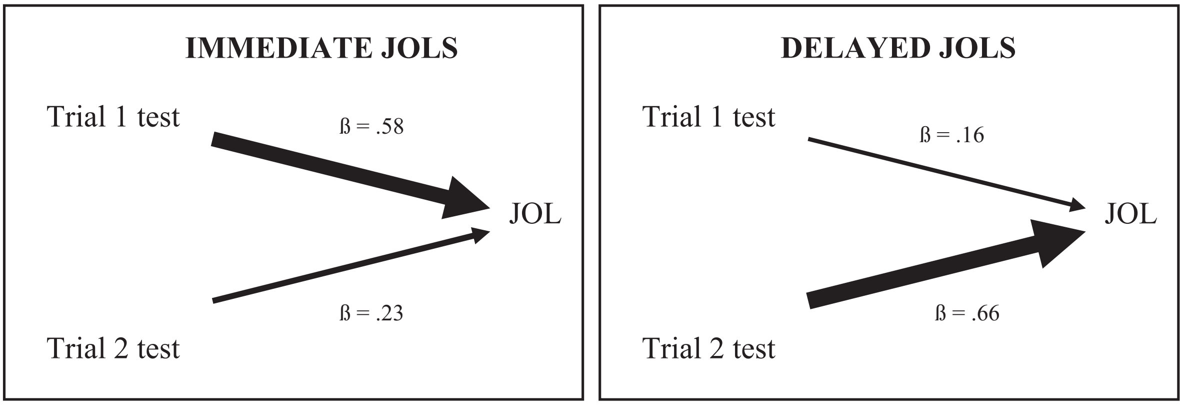

Figure 1. Standardized beta coefficients for regressors Test 1 and Test 2 on Trial 2 judgments of learning (JOLs) in Experiment 1. The left panel shows immediate JOLs; the right panel shows delayed JOLs. The width of the arrows is proportional to the beta coefficients.

Design. The experiment was a 2 (trial: 1 or 2) × 2 (delayed or immediate JOL) within-participants design with 24 word pairs per JOL condition.

Materials and apparatus. The word lists were 48 cue–target pairs made up of concrete nouns that ranged from high (window–pane) to low (tulip–brush) associability, as normed on Columbia undergraduates in a separate study. Mean length of cue and target was 6.24 letters. Each participant was given a distinct random selection of pairs slated for immediate and delayed JOLs. Classification of the pairs into delayed or immediate JOL condition was retained throughout both trials.

Procedure. Participants were instructed that they would be studying 48 cue–target pairs (24 pairs per JOL condition), making JOLs, taking a cued-recall test, and that this procedure would happen twice. JOLs were explained as in Experiment 1. Cue–target presentation and the recall test followed the same procedure as in Experiment 1.

Results

Recall performance means for Trial 1 were .50 and .47 in the immediate and delayed conditions, respectively; for Trial 2 they were .82 and .80. Recall increased from Trial 1 to Trial 2, $F(1, 29) = 291.10$, $MSE = .01$, $p < .001$, $\eta^2 = .91$. Immediate and delayed JOL conditions showed no differences in recall performance, and there was no interaction between trials and condition on recall ($F < 1$).

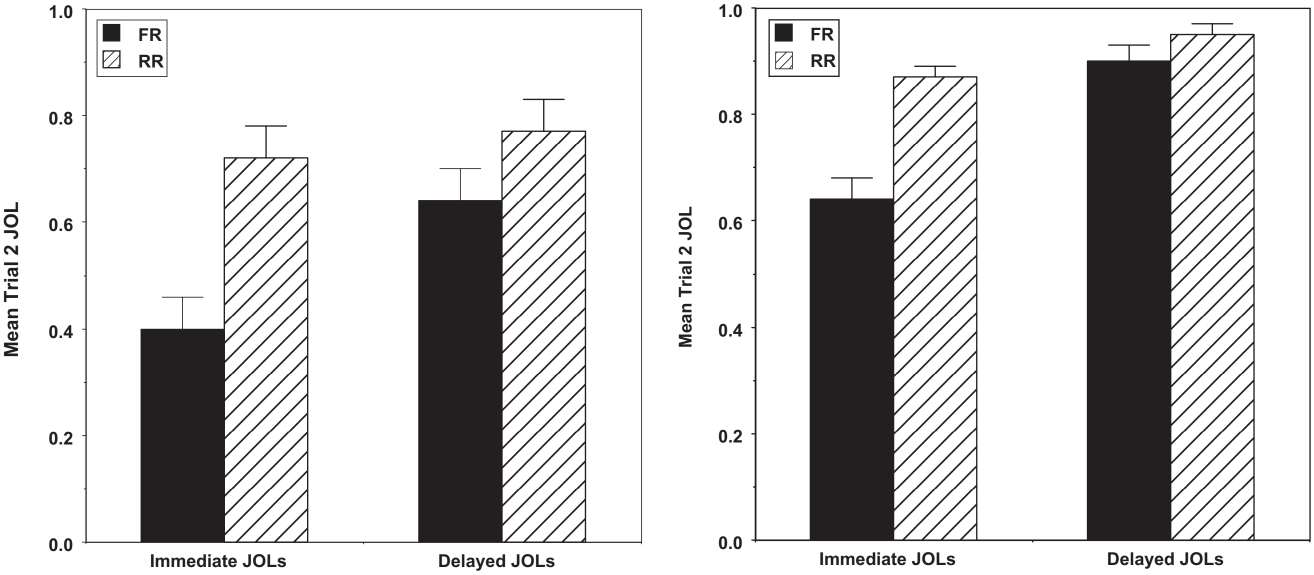

Figure 2. Mean Trial 2 judgments of learning (JOLs) for items forgotten on Test 1 and remembered on Test 2 (FR) and remembered on Test 1 and remembered on Test 2 (RR) in both immediate- and delayed-judgment conditions. The left panel shows FR and RR means for immediate and delayed conditions in Experiment 1, and the right panel shows the same in Experiment 2. Bars indicate standard error.

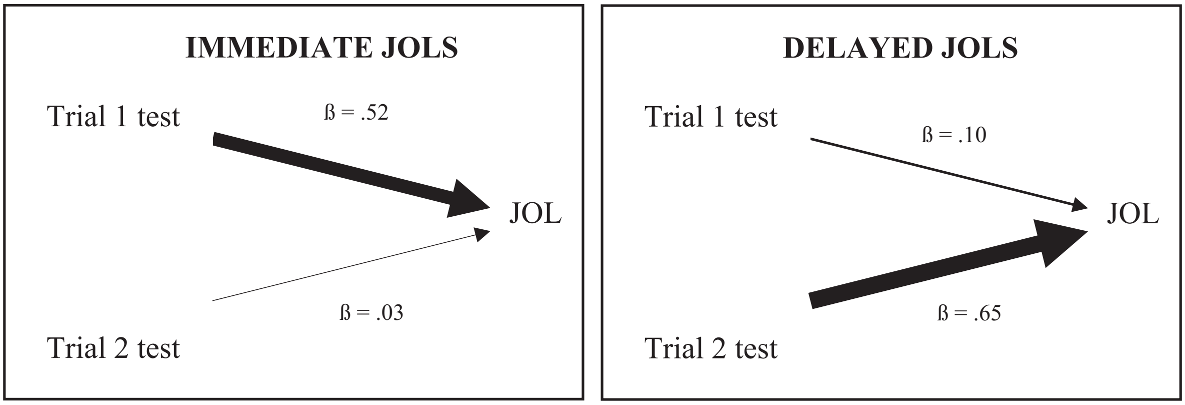

Figure 3. Standardized beta coefficients for regressors Test 1 and Test 2 on Trial 2 judgments of learning (JOLs) in Experiment 2. The left panel shows immediate JOLs; the right panel shows delayed JOLs. The width of the arrows is proportional to the beta coefficients.

Trial 1 immediate JOL mean was .63, and delayed was .55. Trial 2 immediate JOL mean was .75, and delayed was .82. JOLs increased over trials, $F(1, 29) = 61.40$, $MSE = .02$, $p < .001$, $\eta^2 = .68$. There was also a large and significant interaction between trials and condition, $F(1, 29) = 10.18$, $MSE = .02$, $p = .003$, $\eta^2 = .26$. Post hoc tests showed that immediate and delayed JOLs were significantly different on Trial 1, $t(29) = 2.18$, $p = .04$, and on Trial 2, $t(29) = 2.26$, $p = .03$.

We computed calibration to examine whether JOLs were biased. There was an effect of trial, $F(1, 29) = 17.88$, $MSE = .03$, $p < .001$, $\eta^2 = .38$; overconfidence was exhibited on Trial 1, $t(29) = 3.04$, $p = .01$; on Trial 2, calibration was not significantly different from zero ($t < 1$). The Trial × Condition interaction was significant, $F(1, 29) = 9.16$, $MSE = .02$, $p = .01$, $\eta^2 = .24$. Both the immediate ($M = .13$), $t(29) = 2.52$, $p = .02$, and delayed ($M = .08$), $t(29) = 3.45$, $p = .002$, conditions evidenced overconfidence on Trial 1 (with the difference between them failing to reach significance, by a t test). On Trial 2, calibration for the delayed-judgment condition was not significantly different from zero ($M = .03$), $t(29) = 1.15$, $p = .26$, whereas the immediate-judgment condition showed significant underconfidence ($M = –.07$), $t(29) = 2.24$, $p = .03$. The difference between these two conditions was significant on Trial 2, $t(29) = 3.89$, $p = .001$.

As in Experiment 1, we computed an ANOVA using standardized beta coefficients computed from each participant's immediate and delayed simultaneous regression models, regressing each item's Trial 2 JOL on Trial 1 and Trial 2 tests. As shown in Figure 3, results were similar to those of Experiment 1.[3] The Test Trial × JOL Condition interaction was significant, $F(1, 21) = 129.07$, $MSE = .05$, $p < .001$, $\eta^2 = .86$. Post hoc tests demonstrated that Trial 1 and Trial 2 tests were significantly different in the immediate JOL condition, $t(26) = 7.73$, $p < .001$, and in the delayed JOL condition, $t(23) = 7.56$, $p < .001$.

Lastly, we computed the 2 (JOL condition: immediate vs. delayed) × 2 (item status: FR vs. RR) within-participants ANOVA. There was a main effect of test status, $F(1, 29) = 45.61$, $MSE = .01$, $p < .001$, $\eta^2 = .61$; a main effect of condition, $F(1, 29) = 42.02$, $MSE = .02$, $p < .001$, $\eta^2 = .59$; and, critically, a JOL Condition × Item Status interaction, $F(1, 29) = 25.42$, $MSE = .01$, $p < .001$, $\eta^2 = .47$, shown in Figure 2. The difference between RR and FR was four times larger in the immediate condition (.24) than in the delayed condition (.05), $t(29) = 5.04$, $p < .001$. Immediate FR items showed the lowest mean JOLs overall (.64) and were significantly lower than the delayed FR items, $t(29) = 6.46$, $p < .001$.

General Discussion

Both experiments implicate the MPT heuristic after Trial 1 when making immediate JOLs. Simultaneous multiple regression showed that the past test predicts immediate JOLs better than does the upcoming testing. Furthermore, items that were forgotten on the past test but remembered on the next test, in the immediate condition alone, contribute disproportionately to the underconfidence effect. It is with the immediate condition, and not the delayed JOL condition, that the UWP effect is observed.

In the absence of better information (available with delayed JOLs), memory for past performance may be a relatively accurate predictor of future performance. It is interesting to note that it may also be adaptive in allowing the learner to sort out items that need no further study from those that are appropriate for the allocation of additional study time and effort. The MPT heuristic, whereby people base their metacognitions on their remembered past test performance and fail to adequately take into account new learning, appears to be a way to target items that could benefit from additional study time.

Although divergences from calibration have been taken to flag inadequacies in metacognitive control, we suggest that the repercussions of people's underconfidence—because it is based, at least in part, on the MPT heuristic—might have a positive spin. Additional attention and study effort to the just-learned items—items that are fragile and need bolstering—may be an unexpected benefit of what might otherwise be deemed simply another metacognitive error.

[3] Results were also identical to Experiment 1 in an analysis that used ranked JOLs in the regression models for each participant.

References

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral and Decision Processes, 10,* 153–285.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Cohen, J., Cohen, P., West, S., & Aikin, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition, 25,* 691–700.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology, 3,* 552–564.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). Norms for the Toronto Word Pool. *Behavior Research Methods and Instrumentation, 14,* 375–399.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15,* 227–233.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16,* 45–54.

Halff, H. M. (1977). The role of recall opportunities in learning to retrieve. *American Journal of Psychology, 90,* 383–406.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions and memory task performance in adults. *Psychology and Aging, 5,* 215–227.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 22–34.

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

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

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

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

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

LaPorte, R., & Voss, J. F. (1974). Paired-associate acquisition as a function of number of initial nontest trials. *Journal of Experimental Psychology, 103,* 117–123.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 756–766.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123–132.

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.

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

Scheck, P., & Nelson, T. (2005). Lack of pervasiveness of the underconfidence-with-practice-effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134,* 124–128.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect*? Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1258–1266.

Sikström, S., & Jönsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review, 112,* 932–950.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 5,* 315–316.

Thiede, K. W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin and Review, 6,* 662–667.

Received March 17, 2006
Revision received August 16, 2006
Accepted August 18, 2006 ■