Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 58 (2008) 19–34
Journal of Memory and Language
www.elsevier.com/locate/jml

Judgments of learning are influenced by memory for past test ☆

Bridgid Finn *, Janet Metcalfe

Department of Psychology, Columbia University, 406 Schermerhorn Hall, New York, NY 10027, USA

Received 26 July 2006; revision received 9 March 2007
Available online 3 May 2007

Abstract

The Underconfidence with Practice (UPW) effect [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.], found in multi-trial learning, is marked by a pattern of underconfidence accompanied by an increase in resolution between the judgments and test on and after the second trial. We tested whether the memory for past test (MPT) heuristic [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.*] could explain the resolution and calibration effects. To selectively alter Trial 1 test performance, and hence MPT, we manipulated the number of repetitions (Experiment 1) or the study time (Experiment 2) on Trial 1, but then the manipulation was reversed on Trial 2, thereby equating final performance. Despite equivalent Trial 2 recall performance, Trial 2 JOLs reflected the manipulated Trial 1 test performance, providing support for the MPT hypothesis. Follow up experiments tested alternative explanations. We found that people could remember past test and that use of this information would produce both underconfidence and improved resolution. In contrast, neither memory for Trial 1 encoding fluency nor memory for Trial 1 JOLs was able to explain both aspects of the UWP effect. These experiments support the proposal that people use the memory for past test heuristic to make second trial immediate JOLs, and that its use can account for the UWP effect.

Keywords: Judgments of learning; Metacognition; Underconfidence; Memory for past test heuristic

Metacognitive monitoring is thought to be central in guiding people's learning behavior, influencing how much study time people allocate to a particular item, and which items they choose to study (see Nelson, Dunlosky, Graf, & Narens, 1994). Insofar as people's judgments about their learning influence their study efforts, any inaccuracies in the judgments may result in less effective learning. While metacognitive monitoring is fairly accurate in predicting upcoming memory performance there are cases in which metamemory judgments do not appropriately reflect what they are meant to appraise (Benjamin, Bjork, & Schwartz, 1998; Koriat & Bjork, 2005; Zechmeister & Shaughnessy, 1980). Judgments can exhibit systematic biases (Koriat, 1997; Metcalfe, 1998), which may in turn lead to the selection of less efficient study strategies.

A bias that has been accorded much recent attention is the underconfidence with practice (UWP) effect (Kori-

☆ This research was supported by National Institute of Mental Health Grant RO1-MH60637. We thank Nate Kornell and Matt Greene for their help and comments.
* Corresponding author.
E-mail address: bmf2003@columbia.edu (B. Finn).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.03.006

at, Sheffer, & Ma'ayan, 2002). The UWP effect is characterized by judgments of learning (JOLs) that underestimate recall performance on and after Trial 2. This calibration bias occurs in combination with an increase in resolution, which is a measure of the learner's sensitivity for which items will be remembered and which items will be forgotten.

This article tests the memory for past test (MPT) heuristic (Finn & Metcalfe, 2007) as an explanation of the UWP effect. This heuristic states that in the absence of better diagnostic information people may rely on their memory of their performance in the last test in making their JOLs. As will be detailed shortly, the MPT heuristic can account for both the underconfidence and the improvements in resolution that are found with multitrial immediate JOLs, on the second trial and beyond.

A number of studies have ruled out other potential explanations of the UWP effect (Koriat, 1997; Koriat et al., 2002; Serra & Dunlosky, 2005). For example, Serra and Dunlosky (2005) eliminated a retrieval fluency explanation. That explanation proposed that items lacking retrieval fluency, because they had been recalled with difficulty on Trial 1, might be given inappropriately low subsequent JOLs, resulting in selective underconfidence. The JOLs would be inappropriate because items remembered on Trial 1, even with difficulty, are also likely to be recalled on Trial 2 (Benjamin & Bjork, 1996). In contrast to their expectations, Serra and Dunlosky (2005) did not find selective underconfidence for items that lacked retrieval fluency.

Koriat et al. (2002) ruled out three plausible explanations of the UWP effect: the underestimation of performance explanation, the study time allocation explanation and the hard–easy effect explanation. The underestimation of performance explanation proposed that the UWP effect may arise because people underestimate their prior recall performance, and that the underestimation biases the upcoming JOLs. This explanation was rejected because when people were given feedback about how they had performed, which was expected to remedy the underestimation of their earlier recall performance, the UWP effect still occurred. Koriat et al. (2002) also eliminated a study time allocation explanation. Their idea was that most experiments present items to participants at a fixed pace and if participants feel they have not seen a particular item for long enough, they may give it a low JOL. That explanation was discarded because when people were allowed to subjectively control their own study, underconfidence still occurred. Finally, they addressed the hard–easy effect as an explanation of UWP. It has been found that, in general, easy materials give rise to less overconfidence than do difficult materials. It was conceivable then, that as items became easier with learning on each trial, JOLs would become underconfident. This explanation was ruled out because Koriat et al. (2002) showed that both easy and difficult items showed the UWP effect. Thus, while the UWP phenomenon is robust, explanations are lacking.

Recently, Finn and Metcalfe (2007) have proposed that in the absence of more diagnostic information, people may make their Trial 2 immediate JOLs based on their memory for past test (MPT), and that this heuristic may explain the UWP effect. If participants remember having answered correctly on a test, they give that item a high JOL; if they remember having forgotten it on the test, they give it a low JOL. If immediate JOLs rely primarily on the MPT heuristic, then any improvement in recall from trial to trial would tend to be underestimated, insofar as learning has occurred since the test. Reliance on the MPT heuristic should, therefore, result in systematic underconfidence because it reflects past test performance without sufficiently taking into account new learning. Indeed, when the MPT heuristic is not available, as in Trial 1, the typical finding is overconfidence. The standard finding on and after Trial 2 is underconfidence (Koriat et al., 2002).

Finn and Metcalfe (2007) argued that the MPT heuristic was used when other more diagnostic cues were not available, namely, when judgments were made immediately but not at a delay. Delayed judgments, in contrast to immediate judgments are thought to involve an attempt at target retrieval. The success or failure of that attempt is diagnostic information that precludes the need to rely on the MPT heuristic. Alternatively, immediate JOLs, which do not involve a diagnostic retrieval attempt, do have reason to rely on MPT. In support of the MPT heuristic, Finn and Metcalfe (2007) have noted that when judgments are delayed, the UWP effect is either reversed (Dunlosky & Connor, 1997; Koriat & Ma'ayan, 2005; Koriat, Ma'ayan, Sheffer, & Bjork, 2006), absent (Finn & Metcalfe, 2007; Meeter & Nelson, 2003), or extremely truncated (Serra & Dunlosky, 2005). Furthermore, Finn and Metcalfe (2007) showed that past test performance predicted immediate JOLs, but not delayed JOLs. They presented a simultaneous multiple regression analysis, which showed that with immediate JOLs, an item's Trial 1 test performance was a better predictor of its Trial 2 immediate JOL than was its Trial 2 test performance. The opposite pattern emerged with delayed JOLs: an item's Trial 2 test performance better predicted its Trial 2 JOL than did its Trial 1 test performance. They also found, as a reliance on the MPT heuristic would suggest, that items that were forgotten on Trial 1 but remembered on Trial 2 (newly learned items) disproportionately contributed to the UWP effect, but critically, only when the judgments were immediate. They argued that the reliance on information about whether an item was or was not recalled on the prior test, without sufficient consideration of new learning, could account for selective UWP with immediate JOLs.

There are two critical features of the UWP effect that we will take as being the hallmark of the effect: underconfidence and improvements in resolution across trials. Over study–test trials calibration becomes negatively biased toward underconfidence whereby the mean JOL underestimates the mean recall performance. This occurs in conjunction with improvements in resolution, which indicate that people are becoming better at discriminating which items will be remembered and which items will be forgotten. In 15 separate experiments on the UWP effect, using repeated study–test trials both aspects of this pattern emerged, without exception (Finn & Metcalfe, 2007; Koriat, 1997; Koriat, Ma'ayan, & Nussinson, 2006; Koriat et al., 2002). Moreover, as a reliance on MPT would predict, in the two studies investigating the UWP effect in which no test occurred between Trial 1 and Trial 2 (Koriat, 1997; Koriat & Bjork, 2006; Meeter & Nelson, 2003), while judgments showed underconfidence, there was no accompanying improvement in resolution (but see Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002).

It is our conjecture here that use of the MPT heuristic can account both for the differences in calibration and differences in resolution accuracy that define the UWP effect. As Finn and Metcalfe (2007) have demonstrated, reliance on MPT leads to underconfident judgments that are highly correlated to the prior test. The MPT heuristic should also lead to high resolution between predictions and performance, insofar as recall performance remains reasonably stable between tests. Since Trial 1 and Trial 2 recall performance are highly correlated, using prior test performance to predict upcoming performance is a highly diagnostic strategy. On Trial 1 this strategy is not available of course. Gamma correlations, which index resolution, should increase once this cue does become available on Trial 2. This prediction is consistent with the many multi-study–test-trial studies showing that resolution improves after a test (e.g. Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; King, Zechmeister, & Shaughnessy, 1980; Koriat & Bjork, 2006; Lovelace, 1984; Mazzoni, Cornoldi, & Marchitelli, 1990).

The evidence that Finn and Metcalfe (2007) provided was an important start in showing that the UWP effect might be due to reliance on memory for past test. However, it was essentially correlational. To be more convincing, our first goal was to show that an experimental manipulation that altered performance on the prior test, but equated performance on the upcoming test altered the UWP effect itself. Accordingly, we here manipulated Trial 1 test performance, without altering people's eventual test performance on Trial 2. This enabled an observation of the predicted variations in underconfidence as a result of the manipulation.

In these experiments, on Trial 1, to be remembered items within a list received one or many repetitions (in Experiment 1) or more or less presentation time (Experiment 2). On Trial 2, this pattern was reversed such that the items that had received one study repetition, or little study time were now accorded many repetitions or much study time. By the end of Trial 2, each item had received the same total number of repetitions or presentation time, and recall performance was equated. In this way we were able to vary memory performance on Trial 1 test, while equating it on Trial 2.

Our hypothesis was that if Trial 2 JOLs rely on the MPT heuristic, then when performance is matched on Trial 2, JOLs should show differences that reflect outcomes of the prior test. The specific prediction was that items given fewer repetitions or less time on Trial 1, and hence had worse test performance, should also have lower Trial 2 JOLs than items given more repetitions or time on Trial 1. To allow a fair test we also needed to show that performance on Trial 2—the trial about which the judgments were being made—was the same.

Experiment 1

In condition 5–1, cue-target pairs were repeated 5 times on Trial 1 and only once on Trial 2. In condition 1–5, equivalent cue-target pairs were repeated once on Trial 1 and 5 times on Trial 2. Our hypothesis was that if people relied on the MPT heuristic then the Trial 2 JOLs should be lower in the 1–5 condition than in the 5–1 condition, mirroring Trial 1 (but not Trial 2) test performance. If only eventual Trial 2 test performance was important, then the two conditions should exhibit equal JOLs. We used a fully crossed design, in which the number of presentations on Trial 1 (5 or 1) was crossed with the number on Trial 2 (5 or 1).

Method

Participants

The participants were 42 undergraduates at Columbia University and Barnard College. They participated for course credit or cash and were treated in accordance with the APA ethical guidelines.

Design

The experiment consisted of a 2 (trial: 1 or 2) × 2 (presentation repetitions on Trial 1: 1 or 5) × 2 (presentation repetitions on Trial 2: either 1 or 5) within-participants design with 12 word pairs per treatment combination. The four repetition treatment combinations were combined within a single list, for each participant.

Materials

The word lists were 48 cue-target pairs made up of concrete nouns taken from Paivio, Yuille, and Madigan

(1968). Mean word length of both cue and target was 7.34 letters, and no words exceeded 8 letters. For each participant, the computer randomly combined the words into pairs and randomly selected which pairs would be slated for each study/repetition condition.

Procedure

Participants were instructed that they would be learning 48 word pairs and making JOLs. JOLs were explained as judgments of learning based on what they thought were their chances for recalling the second word when given the first word during a memory test that would happen in a few minutes. They were asked to use a scale from 0 to 100%. They were also told that at test they would be given the cue word and would have to type in the target.

On Trial 1, half of the pairs were presented once, and half were presented 5 times. Each presentation was 3s. On Trial 2 half of the pairs that had received 1 presentation received 1 additional presentation and the other half received 5 presentations. Similarly, half of the pairs that had received 5 presentations on Trial 1 received 1 repetition on Trial 2 while the other half received 5 presentations resulting in 4 presentation conditions: 1–1, 1–5, 5–1 and 5–5.

After pairs slated for five presentations were shown four times the remaining study presentation for that condition, and the one-repetition pairs were randomly shuffled by the computer and given their final or only presentation. This final presentation was immediately followed by a JOL in which the cue was presented and participants were asked to type in a JOL ranging from 0 to 100%.

After the first trial, participants were tested on all word pairs. The cue was presented and they were asked to type in the target. There were no restrictions on the amount of time they could spend on the test. They were not given feedback. They then proceeded to Trial 2 study, Trial 2 JOLs and Trial 2 test.

Results

Recall performance

We were not particularly interested in the 1–1 and 5–5 conditions, except insofar as they showed that the repetition manipulation was behaving lawfully, which it was. Thus, across all experiments, only the analyses that are relevant to our conditions of interest (1–5, 5–1) will be reported. As expected, there was an effect of trials, such that recall performance improved from Trial 1 to Trial 2, $F(1,41) = 231.55$, $MSE = .02$, $p < .001$, $\eta_p^2 = .85$ (effect size is reported as partial eta squared, η_p^2). The trial by repetition condition interaction was significant, $F(1,41) = 102.53$, $MSE = .02$, $p < .001$, $\eta_p^2 = .71$. Planned comparisons revealed that although there was a large difference in performance between the 1–5 and the 5–1 group on Trial 1 ($M = .11$, $SE = .02$ vs. $M = .49$, $SE = .04$), respectively, for a difference of .38, $t(41) = 10.43$, $p < .001$, $CI_{.95} = .31, .46$ (95% confidence intervals are used throughout), there was no difference between these two groups on Trial 2 performance, ($t < 1$, $p > .05$). These recall performance data for conditions 1–5 and 5–1, shown in the left panel of Fig. 1, put us in position to evaluate the hypothesis.

JOLs

Our main interest was in the difference between the 1–5 and 5–1 conditions on the Trial 2 JOLs. As can be seen from Fig. 2, left panel, Trial 2 JOLs were significantly lower for the 1–5 condition ($M = .53$, $SE = .04$) than for the 5–1 ($M = .60$, $SE = .04$) condition, by a difference of .07, $t(41) = 2.99$, $p < .01$, $CI_{.95} = .02, .12$. These data offer support for the MPT hypothesis. The Trial 1 JOL means were also different, .41 and .62, respectively, as expected.

Calibration

The calibration score for each participant was calculated by subtracting mean recall performance from mean

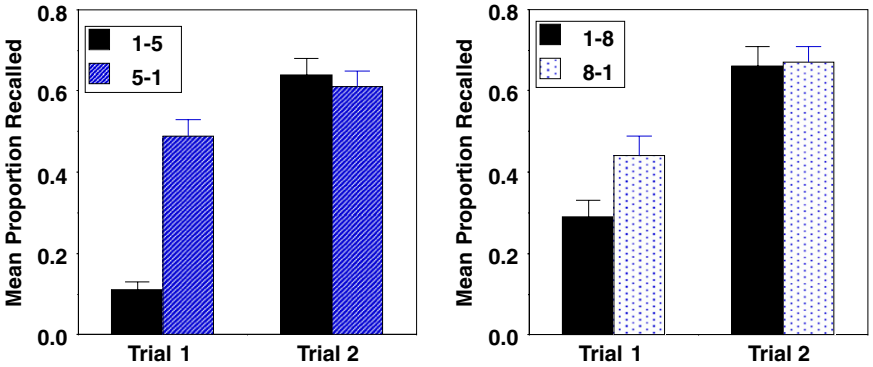

Fig. 1. (Left) Mean proportion recalled in Trial 1 and Trial 2 for repetition conditions 1–5 and 5–1 in Experiment 1. Error bars depict standard error of the means. (Right) Mean proportion recalled in Trial 1 and Trial 2 for time conditions 1–8 and 8–1 in Experiment 2. Error bars depict standard error of the means.

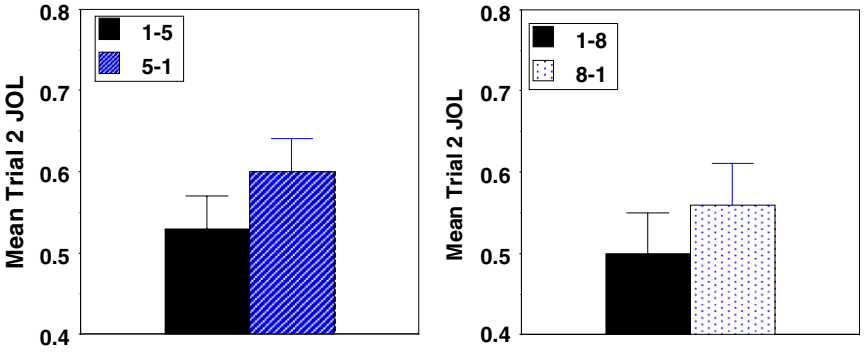

Fig. 2. (Left) Mean Trial 2 JOLs for repetition conditions 1–5 and 5–1 in Experiment 1. Error bars depict standard error of the means. (Right) Mean Trial 2 JOLs for time conditions 1–8 and 8–1 in Experiment 2. Error bars depict standard error of the means.

JOL within each condition. Under and overconfidence were obtained if that score was significantly negative, in the former case, or positive, in the latter. Our interest was in Trial 2 underconfidence. The 1–5 and 5–1 conditions showed a significant .09 difference in Trial 2 underconfidence, $t(41) = 2.98$, $p = .01$, $CI_{.95} = .03, .16$. The 1–5 condition, the condition in which more items were incorrect on Trial 1 test, was significantly more underconfident ($M = -.10$, $SE = .04$, $t(41) = 2.63$, $p = .01$) than the 5–1 condition ($M = -.01$, $SE = .04$), which was not significantly different from zero ($t < 1$, $p > .05$).

JOLs conditionalized on Trial 1 and Trial 2 recall performance

Next, we used an analysis developed in Finn and Metcalfe (2007) to examine whether unrecalled items on Test 1 that were subsequently remembered on Test 2 contributed disproportionately to the UWP effect. If the MPT heuristic were being used, then items forgotten on Trial 1 but remembered on Trial 2 (FR) should have lower JOLs than items remembered on both trials (RR). According to the MPT heuristic people are remembering their item specific performance on the prior test and so we did not expect to find differences between the 1–5 and 5–1 conditions in either the FR or RR category. (We were sometimes unable to report a FR or RR participant mean for a particular repetition condition, because some got everything right or everything wrong in one condition and the statistic could not be computed. Thus, degrees of freedom listed for the conditionalized analyses may differ from those expected from the total number of participants used in the experiment.)

There were no significant differences between the 1–5 and 5–1 conditions for the FR items, $t(35) = 1.58$, $p > .05$ or the RR items, $t(25) = 1.10$, $p > .05$. As can be seen in Fig. 3, left, the mean JOL for the FR items ($M = .57$, $SE = .04$) was significantly lower than JOLs given to RR items ($M = .81$, $SE = .03$), for a difference of .24, $t(39) = 8.45$, $p < .001$, $CI_{.95} = .18, .29$. As predicted by use of the MPT heuristic, FR items were given lower JOLs. This result is consistent with the idea that people were remembering their item-by-item perfor-

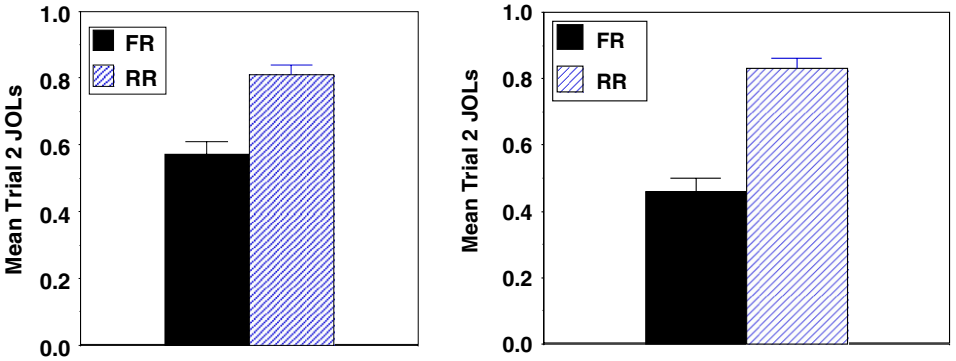

Fig. 3. (Left) Mean Trial 2 JOLs for items forgotten on Test 1 and remembered on Test 2 (FR) and remembered on Test 1 and remembered on Test 2 (RR) in Experiment 1. Error bars depict standard errors of the means. (Right) Mean Trial 2 JOLs for items forgotten on Test 1 and remembered on Test 2 (FR) and remembered on Test 1 and remembered on Test 2 (RR) in Experiment 2. Error bars depict standard errors of the means.

mance on the Trial 1 test, which affected their Trial 2 JOLs.

If people were relying on MPT alone and completely discounting new learning then JOLs for the FR items should have been the same as the items forgotten on both trials (FF). On Trial 1 JOLs for the FF and FR items were not different ($M = .46$, $SE = .04$, $M = .43$, $SE = .03$, respectively, $t(38) = 1.22$, $p > .05$), but by Trial 2, the JOLs were different ($M = .41$, $SE = .04$, $M = .55$, $SE = .04$, respectively, for a difference of .14, $t(38) = 5.92$, $p < .001$, $CI_{.95} = .10, .20$). The fact that Trial 2 JOLs were different for the FR and FF items indicated that something more than memory for past test was contributing to the Trial 2 JOL. It may be that a diluted evaluation of new learning in conjunction with memory for past test contributes to the judgment. So, prior test performance does not appear to be the only influence on the judgment.

Gamma correlations between JOLs and test

We first report the gamma correlations between the JOLs on Trial 1 and 2 with Test 1 and 2, respectively, for 18 participants using a 2 (trials: 1 and 2) × 2 (repetition condition: 1 repetition Trial 1 and 5 repetitions Trial 2, or 5 repetitions Trial 1 and 1 repetition Trial 2) ANOVA. Gamma correlations for the 1–5 and 5–1 conditions were not significantly different on either Trial 1 or Trial 2 (all $t < 1$, all $p > .05$). Gammas improved from Trial 1 ($M = .46$, $SE = .13$) to Trial 2 ($M = .69$, $SE = .08$, $F(1,17) = 6.06$, $MSE = .15$, $p = .03$, $\eta_p^2 = .26$). This increase in relative accuracy, or resolution, from Trial 1 to Trial 2, is consistent with the other studies on the UWP effect and with the MPT predictions.

The MPT hypothesis indicates that people rely on their memory for whether or not they got an item correct on the previous test, in making their second trial JOLs. This hypothesis indicates that the correlations relating second trial JOLs to whether items were or were not correct on the first trial test should be high (though not necessarily different between conditions). Therefore, we computed these 'backward' gammas as well for the two conditions of focal interest. The 1–5 and 5–1 repetition conditions were not significantly different from one another, $t(17) = 1.39$, $p > .05$. Consistent with what one might expect from the MPT hypothesis, the backward gamma mean was higher ($M = .85$, $SE = .06$) than the gammas between the second trial JOLs and the second trial tests, ($M = .69$, $SE = .07$, $F(1,17) = 10.00$, $MSE = .04$, $p = .01$, $\eta_p^2 = .37$). (No other simple effects or interactions were significant. Note that we did not use a simultaneous multiple regression here, as had been done in Finn and Metcalfe (2007), because the level of performance, which was systematically manipulated on the first test here, affects this parametric assessment of the correlations. Therefore it was inappropriate in the present context. Gamma correlations, however, see Nelson, 1984, are robust under different levels of performance, and hence are more interpretable in the present context).

Discussion

These results provide support for the hypothesis that people rely on the MPT heuristic in making immediate JOLs on Trial 2 and that doing so gives rise to both underconfidence and gamma improvements. First, when test performance was low on Trial 1, Trial 2 JOLs were correspondingly low. When it was high, they were high. This difference was found despite the fact that performance was equated on the upcoming test in the 1–5 and 5–1 conditions, the test about which people were supposed to be making predictions. We also found that across conditions, JOLs were disproportionately low for items that had been forgotten on a previous test, demonstrating that JOLs were driven by memory for the item specific prior test performance. There was also a sizable correlation between whether items were right or wrong on Trial 1 test, and their item-by-item Trial 2 JOLs. This 'backwards' gamma correlation was significantly larger than was the gamma correlation between Trial 2 JOLs and Trial 2 test performance. Finally, we found a large improvement in resolution over trials.

We sought to determine the reliability our findings, as well as to establish some generality, by performing a second experiment with some modifications in the methodology, but targeting the same question. Experiment 2 was designed to serve as a conceptual replication of Experiment 1, but using a different manipulation to alter the level of Test 1 performance.

Experiment 2

In Experiment 2, we changed the manipulation that altered Trial 1 test performance to presentation rate rather than number of repetitions. In a manner similar to that of additional study presentations, additional presentation time also facilitates recall (Kintsch, 1970; Murdock, 1974). Thus, our basic hypothesis was unaltered, but our method for testing it differed. On Trial 1, items were presented for either 1s or 8s. We expected test performance on Trial 1 to reflect those time differences. On Trial 2 those times were exchanged. We expected that performance would be matched on the Trial 2 test since total presentation time was equivalent, but that JOLs would reflect prior test performance, as in Experiment 1.

Method

Twenty-eight undergraduates at Columbia University and Barnard College participated for course credit

or cash and were treated in accordance with APA ethical guidelines. Four additional participants were not analyzed beyond finding that they had everything incorrect on the first trial.

The experiment consisted of a 2 (trial: 1 or 2) × 2 (time condition: 1 s Trial 1 and 8 s Trial 2, or 8 s Trial 1 and 1 s Trial 2) within-participants design with 24 word pairs per treatment combination, yielding 48 pairs per list. Experiment 2 used the same materials and procedure as in Experiment 1 except that a presentation time manipulation was used instead of a repetition manipulation. On Trial 1, half of the pairs were presented for 1s, half were presented for 8 s. Presentation of 1 or 8 s was randomized. On Trial 2 the pairs that had received a 1 s presentation received an 8 s presentation and the pairs that had received an 8 s presentation on Trial 1 received a 1 s presentation making up 2 presentation conditions: 1–8 and 8–1.

Results

Recall performance

Mean recall performance improved from Trial 1 to Trial 2, $F(1,27) = 406.64$, $MSE = .01$, $p < .001$, $\eta_p^2 = .94$. There was an effect of time condition, $F(1,27) = 10.85$, $MSE = .02$, $p = .003$, $\eta_p^2 = .29$, such that over the two trials the mean recall for the 8–1 time condition was higher (.56) than was the 1–8 time condition (.47). The interaction between time condition and trial was significant, $F(1,27) = 29.74$, $MSE = .01$, $p < .001$, $\eta_p^2 = .52$. There was a large difference in recall performance between the 1–8 and the 8–1 group on Trial 1 ($M = .29$, $SE = .05$ vs. $M = .44$, $SE = .08$, respectively, for a difference of .15, $t(27) = 5.40$, $p < .001$, $CI_{.95} = .10, .21$), but no difference between these two groups on Trial 2 (.66 vs. .67, respectively, $t < 1$, $p > .05$). See Fig. 1, right panel.

JOLs

A planned comparison showed that Trial 2 JOLs were higher for the 8–1 ($M = .56$, $SE = .05$) time condition than for the 1–8 ($M = .50$, $SE = .05$) time condition, for a difference of .06, ($t(27) = 2.50$, $p = .02$, $CI_{.95} = .01, .10$), as is shown in the right panel of Fig. 2. On Trial 1 the mean JOLs were: 1–8: .33 and 8–1: .47.

Calibration

Both the 1–8 and 8–1 conditions were underconfident on Trial 2, but the 1–8 time condition was significantly more so ($M = -.16$, $SE = .03$, $t(27) = 6.24$, $p < .001$) than the 8–1 condition ($M = -.12$, $SE = .02$, $t(27) = 5.27$, $p < .001$) for a difference of .04, $t(27) = 2.21$, $p = .04$, $CI_{.95} = .01, .08$. No other main effects or interactions were significant.

JOLs conditionalized on Trial 1 and Trial 2 recall performance

There were no significant differences between the 1–8 and 8–1 conditions for the FR items, $t(35) = 1.58$, $p > .05$, or the RR items, $t(25) = 1.10$, $p > .05$. The mean JOL for the FR items ($M = .46$, $SE = .04$) was significantly lower than JOLs given to RR items ($M = .83$, $SE = .03$), for a difference of .37, $t(27) = 7.84$, $p < .001$, $CI_{.95} = .27, .46$, shown in Fig. 3 right. This result was consistent with the results reported in Experiment 1 and with the idea that Trial 2 JOLs were affected by people's memory of item specific performance on the Trial 1 test.

A comparison of the JOLs given to the FF and FR items revealed that they were not significantly different on Trial 1, $t(25) = 1.24$, $p > .05$, but they did differ on Trial 2, ($M = .34$, $SE = .05$, $M = .44$, $SE = .04$, respectively, for a difference of .10, $t(25) = 4.40$, $p < .001$, $CI_{.95} = .06, .16$), which again suggested that something in addition to memory for past test was contributing to the Trial 2 JOL.

Gamma correlations

Gammas improved from Trial 1 ($M = .27$, $SE = .09$) to Trial 2 ($M = .59$, $SE = .07$). No other effects were significant. The 'backward' gamma correlations between Trial 2 JOLs and Trial 1 test were .80 ($SE = .08$) for the 1–8 condition, and .91 ($SE = .03$) for the 8–1 condition, which showed a .11 significant difference from one another, $t(25) = 2.10$, $p < .05$, $CI_{.95} = .01, .27$. We contrasted these backward gammas with those between Trial 2 JOLs and Trial 2 test. As in Experiment 1, there was a main effect, $F(1,20) = 41.65$, $MSE = .04$, $p < .001$, $\eta_p^2 = .68$, such that the backward gamma correlations ($M = .86$, $SE = .05$) were significantly higher than were the forward gammas ($M = .59$, $SE = .07$). No other effects were significant.

Discussion

Experiment 2 replicated the findings of Experiment 1 using a presentation time manipulation. As in Experiment 1, prior test performance had consequences for the Trial 2 JOLs. Items in the 1–8 condition, which had shown worse Trial 1 test performance than items in the 8–1 condition also showed lower Trial 2 JOLs. This pattern emerged despite equivalent Trial 2 test performance between the 1–8 and 8–1 conditions.

Experiment 3

Experiment 3 tested an implication of the MPT explanation of the UWP effect. We examined whether people could explicitly remember their test performance on the previous trial at the time they would be making

their JOLs and if so, whether this would result in both underconfidence and improved resolution. The literature suggests that people can remember their prior test performance. For example, Gardiner and Klee (1976) demonstrated that people have very good memory for their performance on a prior test. Recently, in support of the MPT heuristic, Dunlosky and Serra (2006) reported that when people were asked about how they made their Trial 2 JOLs, they reported an explicit use of memory for their prior test performance. Others have shown that a test can change encoding strategies on a subsequent trial (Dunlosky & Hertzog, 2000; Gardiner, Passmore, Herriot, & Klee, 1977; Halff, 1977; LaPorte & Voss, 1974), and that recall performance on the prior trial is strongly correlated with JOL ratings on a subsequent trial (Finn & Metcalfe, 2007; Hertzog, Dixon, & Hultsch, 1990; King et al., 1980; Lovelace, 1984; Thiede, 1999). Lately, Finn and Metcalfe (2007) showed that an item's Trial 1 test performance is a better predictor of its Trial 2 JOL, than is its Trial 2 test performance.

In the experiment, people were given two study–test trials, and were asked on the first trial to make immediate JOLs. On the second trial, they either made immediate JOLs on the items, or they gave a judgment of whether they had correctly recalled the target item corresponding to the cue on the previous test trial. This was done as a within-participants design. Asking people to make judgments of whether or not they recalled each item under the same conditions in which they would normally make JOLs had two advantages. First, it enabled us to investigate the accuracy of their memory for their recall performance at the time of JOL. Secondly, it allowed us to assess whether the MPT judgment would give rise to underconfidence and enhanced resolution on Trial 2.

Method

Participants, design, materials and procedure

The participants were 25 undergraduates at Columbia University and Barnard College. They participated for course credit or cash.

The experiment consisted of a 2 (trial: 1 or 2) × 2 (judgment type: MPT or JOL) within-participants design with 24 word pairs per judgment type, yielding 48 pairs per list. The word lists were 48 cue-target pairs made up of concrete nouns taken from Paivio et al. (1968). The mean cue length was 7.00 letters and the target mean was 7.69 letters. Pairs were randomly combined and selected for either a Trial 2 memory judgment or immediate JOL.

During Trial 1, word pairs were presented for 3 s of study followed by an immediate JOL. Participants were tested on Trial 1, then the second trial began. Immediately after each cue-target pair was presented, the cue stayed on, and participants made either an immediate JOL or a MPT judgment. The type of judgment was blocked, such that an individual participant made MPT judgments (or JOLs) for the first half of the list, then switched to making the other kind of judgment in a block, for the second half of the list. Whether MPT judgments or JOLs came first was counterbalanced over participants. Note that each participant made both types of judgments. Judgment order did not show any effect in any analysis and will not be discussed further. Trial 2 JOL instructions were unchanged from those used on Trial 1. They were told in the MPT condition that they would not be making JOLs but rather would be asked for MPT judgments about or whether they had gotten the target for the presented cue correct on the test that they had just taken by pressing a key that said 'yes' or 'no'. We coded MPT 'yes' as 100 and 'no' as 0, in the subsequent analyses. The cues were re-randomized and presented for the Trial 2 recall test, in which participants were shown each cue and asked to type in the target. As in Experiment 1, test responses were self-paced.

Results

Was memory for Trial 1 test performance accurate?

To measure the accuracy of the memory judgments at the time that JOLs were made, we evaluated the conditional probability of saying 'remembered' when an item was actually remembered, and that was .94. In contrast the probability of falsely calling an item that was unrecalled on Trial 1, recalled, was only .03. People exhibited very good memory for what they had recalled on the previous trial. A gamma correlation was computed between the MPT and the recall performance. If people had accurate memories of their Trial 1 test performance this correlation should be high. The gamma correlation was .99 ($SE = .002$) and indicated very accurate memory for performance on the prior trial.

Would use of MPT result in Trial 2 underconfidence?

On Trial 2, the MPT judgments ($M = -.22$, $SE = .03$, $t(24) = 8.83$, $p < .001$) showed the same significant underconfidence ($M = .15$, $SE = .04$, $t(24) = 3.98$, $p = .001$), as the Trial 2 JOLs, $t(24) = 1.76$, $p > .05$. This comparison indicated that reliance on judgments of past

Table 1
Comparison of MPT, T_1EF and T_1J judgments as sources of the UWP effect

	Judgment Accuracy	Under-confidence Overall	Under-confidence 1-5 > 1-5	Resolution improvements
MPT	✓	✓	✓	✓
T_1EF	✗	✗	✗	?
T_1J	✓	✓	✗	✗

memory performance would indeed result in underconfidence of roughly the same magnitude as observed in the JOLs.

Would MPT lead to improvements in resolution by Trial 2?

The first trial gamma correlations were not different for the MPT and JOL conditions, $t(21) = 1.90$, $p > .05$, as was expected since people made JOLs in Trial 1 in both conditions. On Trial 2, where the judgments were different, the MPT gamma correlations were very high ($M = .94$, $SE = .03$), and were better, in fact, than were the Trial 2 JOL gamma correlations, ($M = .63$, $SE = .07$), for a difference of .31, $t(21) = 3.83$, $p = .001$, $CI_{.95} = .14, .48$.[1] Thus, use of MPT would produce an increase in resolution.[2]

Discussion

The data from Experiment 3 indicates that people have very good memory for what they just recalled at the time of making their JOLs. If they did use that information, they would both be underconfident, and would show gamma improvements over trials. Thus, this experiment, in combination with the results of the first two experiments provides converging evidence that people use the MPT heuristic to make Trial 2 JOLs.

While the evidence from Experiments 1, 2 and 3 provide support for the idea that the MPT heuristic can account for the UWP effect, in the final two experiments we investigated two other possibilities. The fact that differences in Trial 1 test were reflected in Trial 2 JOLs, in Experiments 1 and 2, suggests—as we have argued—that the Trial 1 test differences were crucial. Even so, it might be possible that other Trial 1 differences contributed to the effects we observed. One possibility—investigated in Experiment 4—was that people, instead, used Trial 1 encoding fluency (T_1EF) to make their Trial 2 JOLs. Encoding fluency is a measure of how easily a pair was processed or learned during study (Begg et al., 1989) and has been shown to influence immediate JOLs independently of recall (Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat & Ma'ayan, 2005; Mazzoni & Nelson, 1995; but see Dunlosky & Nelson, 1994). Encoding fluency might have been different for the 1–5 and 5–1 conditions at the end of Trial 1, and if it was, then this discrepancy might have led to the difference in Trial 2 JOL bias shown in Experiments 1 and 2.

A second possibility, tested in Experiment 5, was that people might have been using their memory for Trial 1 JOLs (which we will designate: T_1J) as the basis for their Trial 2 JOLs. Trial 1 JOLs are usually lower than those on Trial 2, hence underconfidence might, plausibly, have arisen from this source. Furthermore, as was shown in both Experiments 1 and 2 Trial 1 JOLs were lower for the 1–5 condition, the same condition that had shown more Trial 2 underconfidence. Thus, it might have been that remembered Trial 1 JOLs rather than memory for past test was at the heart of the underconfidence with practice effect.

Insofar as T_1EF, or T_1J could be alternatives that could potentially qualify to explain the UWP effect each would need to effectively act as stand in for the Trial 2 JOL. We held them to the same standard as the MPT heuristic, namely that the source of information would first have to be available to the participant at the time they made their Trial 2 JOLs, (i.e., people would have to accurately remember their performance on that variable on the prior trial)[3] . Its use would need to show the 5–1 (or study time) bias that had been demonstrated in the first two experiments. Its use would need to produce underconfidence, (i.e., it would need to show a mean value that was lower than the mean Trial 2 recall). Finally, it would have to produce increased resolution, (i.e. show the distinctively high gamma correlations that are typically found between Trial 2 JOLs and Trial 2 test).

Experiment 4

In Experiment 4 we investigated the Trial 1 encoding fluency hypothesis (T_1EF). Perhaps the reason for finding lower Trial 2 JOLs in Experiment 1 and 2 for items that had been presented for only one repetition or less time on Trial 1, was that items presented once, in comparison to items presented five times, experienced less fluent encoding on Trial 1. Participants may have been remembering this processing difference when making their Trial 2 JOLs. We examined whether people were able to remember how successful their Trial 1 encoding was for each item, and if so, whether that factor would result in a difference between the 1–5 and 5–1 conditions, whether overall underconfidence would result, and whether the use of Trial 1 encoding fluency would result in increased resolution on Trial 2.

[1] There was no significant difference ($t < 1$, $p > .05$) between the JOL and MPT gammas when recoded binary JOLs (JOLs between 0 and 50 were recoded as 0, and JOLs between 51 and 100 were recoded as 1) were used to compute the JOL-Trial 2 test gamma correlation.

[2] See Table 1 for a summary of the results of Experiments 3 as well as of analogous results for Experiments 4 and 5.

[3] While it is possible that people might use implicit information to make their JOLs, given that in the case of the MPT heuristic the information was explicit, we assumed that the evidence favoring a particular heuristic was more compelling if people were able to access that information.

As in Experiment 1, we had had 4 repetition conditions (1–5, 5–1, 1–1, 5–5, corresponding to the number of repetitions on Trial 1 and Trial 2, respectively). We attempted to isolate memory for Trial 1 encoding fluency as the only potential source of Trial 1 bias by eliminating the Trial 1 JOL and the Trial 1 test. On Trial 2 we asked people to make T_1EF judgments about how well they had encoded each item, by its last Trial 1 presentation. If people were able to remember encoding differences on Trial 1, then they should give higher T_1EF judgments to items in the 5–1 repetition condition than to items in the 1–5 condition. This pattern would suggest that Trial 1 encoding fluency could be used in making Trial 2 JOLs. If no T_1EF judgment differences were found for the 1–5 and 5–1 conditions it would indicate that Trial 1 encoding fluency information was not available while making Trial 2 JOLs.

Method

The participants were 24 undergraduates at Columbia University and Barnard College who participated for course credit or cash and were treated in accordance with the APA ethical guidelines. The design and materials were identical to those in Experiment 1.

The procedure was also identical to Experiment 1 except for the following: On Trial 1 there was no JOL or test phase. In between the first and second study trial participants read a story for about 3 min, which corresponded to the duration of the test in Experiment 1. At the start of the second study trial participants were told that during the current study trial they would be asked to make an encoding fluency judgment. An encoding fluency judgment was explained as a judgment of how fluent their encoding of a pair was the last time it was presented on Trial 1. The instructions included descriptions from the literature about what encoding fluency meant: "Some people have talked about encoding fluency as the ease in learning or processing the word pairs. Others have described it as the ease or speed in forming an image that links the two words." They were asked to use a scale from 0 to 100%. Following the written presentation of the instructions, participants heard instructions verbally to emphasize what the encoding fluency judgment entailed. Finally, after the written and verbal instructions were administered, participants were asked to write down what they thought they should be making a judgment about. Three independent judges reviewed the responses that participants gave about their judgment task. (There was unanimous consent that 8 participants did not respond appropriately. Their data were excluded from all analyses.) Then the items were presented. On Trial 2, the presentation of each item was immediately followed by a T_1EF judgment. A test followed the second trial.

Results

As in previous experiments Trial 2 recall performance for the 1–5 ($M = .58$, $SE = .05$) and 5–1($M = .63$, $SE = .05$) conditions was not significantly different, $t(23) = 1.79$, $p > .05$, and was at about the same level as in previous experiments.

Was memory for Trial 1 encoding fluency (T_1EF) accurate, as reflected in a difference that could account for the 5–1 bias?

We conducted planned comparisons with the 1–5 and 5–1 conditions for the T_1EF judgment. The mean 1–5 ($M = .58$, $SE = .04$) and 5–1 ($M = .56$, $SE = .03$) T_1EF judgments were not different from one another, $t < 1$, $p > .05$. By Trial 2, people were not able to distinguish Trial 1 encoding fluency differences for the 1–5 and 5–1 conditions. This indicated that accurate memory for Trial 1 encoding fluency was not available at the time that they would be making their Trial 2 JOLs. If people had been able to remember T_1EF selectively, then the 5–1 and 5–5 conditions should have been the same. Similarly the 1–5 and 1–1 conditions should have been the same. In contrast, the encoding fluency judgment for the 5–5 condition ($M = .70$, $SE = .03$) was significantly higher than the 5–1 condition ($M = .59$, $SE = .03$), by a difference of .11, $t(23) = 5.33$, $p < .001$, $CI_{.95} = .08, .17$. The 1–1 condition ($M = .31$, $SE = .06$) was significantly lower than the 1–5 condition, by a difference of .27, $t(23) = 7.53$, $p < .001$, $CI_{.95} = .20, .36$. These comparisons provided a secondary index that people could not remember their earlier encoding experiences.

The criterion of accurate memory for earlier encoding fluency differences was not met. Thus, further evaluations of the T_1EF judgment were not considered especially relevant. However, we report them for consistency.

Would use of T_1EF result in Trial 2 underconfidence?

Neither the 1–5 or 5–1 T_1EF judgments were significantly underconfident (all $ts < 1$, all $ps > .05$), nor were the two conditions significantly different from one another ($t = 1.54$, $p > .05$), thus failing to show the pattern found consistently in Experiments 1 and 2. Assuming the 1–5 items experienced less fluent encoding on Trial 1 than the 5–1 items, they should have evidenced greater underconfidence on Trial 2 if memory for Trial 1 encoding fluency was the primary source of the UWP effect.

Would T_1EF lead to improvements in resolution by Trial 2?

In this experiment, because we did not have a Trial 1 test, we could not compute Trial 1 gamma correlations or directly assess an increase in resolution. To address this question, we compared the magnitude of gamma

correlations between the T_1EF judgments and the Trial 2 test with both the Trial 1 JOL—test gamma and the Trial 2 JOL—test gamma in Experiments 1, 2 and 3. We wanted to know whether the T_1EF judgments approximated the initial Trial 1 gammas, or were more similar to the improved Trial 2 gammas. We collapsed across the 1–5 and 5–1 repetition conditions since in all experiments the gamma correlations for the 1–5 and 5–1 conditions were not different from one another. The T_1EF gamma ($M = .49$, $SE = .07$) was not significantly different from either the Trial 1 ($M = .41$, $SE = .04$) or the Trial 2 ($M = .62$, $SE = .03$, $t = 1.83$, $p > .05$) JOL—test gamma. This pattern of results left unclear whether making T_1EF judgments would produce improvements in resolution by Trial 2.

Discussion

T_1EF judgments were not accurate, nor could T_1EF produce underconfidence, demonstrating that memory for Trial 1 encoding fluency could not produce the UWP effect. We are not claiming that encoding fluency is irrelevant when making JOLs. Rather, we suggest that this pattern of data shows that people are not able to explicitly remember their encoding experiences from an earlier trial, and allows us to rule out remembered Trial 1 encoding fluency as a source of the UWP effect. It is quite possible that on Trial 2 people use their current encoding experience to make their JOLs—indeed, this may be why the JOLs for the FF and FR items in Experiments 1 and 2 were not the same. That JOLs were higher for FR items indicates that Trial 2 JOLs are based on something other than just remembered past test, otherwise they would have been identical. Encoding fluency or some other assessment of new learning may contribute to the judgment. However, remembered Trial 1 encoding fluency does not produce underconfidence, which at the limit, seems critical for an explanation of UWP.

Experiment 5a

In Experiment 5a, we investigated the memory for Trial 1 JOLs hypothesis. It was possible that Trial 2 JOLs were lower for items that had been given one presentation or less time on Trial 1 because people were remembering their Trial 1 JOL when making their Trial 2 JOLs. This might have resulted in a downward bias of the 1–5 condition, since in both Experiments 1 and 2, Trial 1 JOLs were lower in the 1–5 than in the 5–1 condition. To investigate this possibility further we appraised the gamma correlations between the Trial 1 and Trial 2 JOLs in our earlier data. If the Trial 2 JOLs were based on the Trial 1 JOLs this correlation should have been close to 1. On the contrary, we found that the Trial 1 JOL—Trial 2 JOL gamma correlation was quite low in Experiment 1 ($M = .31$, $SE = .05$), Experiment 2, ($M = .27$, $SE = .04$) and Experiment 3 ($M = .35$, $SE = .04$). We also assessed the gamma correlation between the Trial 1 JOLs and the Trial 2 test. These should have been quite high but were not ($M = .17$, $SE = .06$ in Experiment 1, $M = .23$, $SE = .07$ in Experiment 2, $M = .35$, $SE = .06$ in Experiment 3). It therefore seemed unlikely that Trial 2 JOLs were made using memory for the prior JOLs.

However suggestive these results, in Experiment 5a we tested whether people were able to explicitly remember the JOL they had given each pair in Trial 1, whether this potential source of Trial 2 JOLs could account for the 5–1 difference, whether it would produce underconfidence in general, and whether it would produce an increase in resolution.

Method

The participants were 24 undergraduates at Columbia University and Barnard College. The experimental materials, design and procedure were exactly the same as in Experiment 4 except for the following: During Trial 1 people made a JOL immediately after the final or only presentation of each item, as in Experiment 1. At the beginning of Trial 2 participants were told that during the current study trial they would be asked to make a new kind of judgment about what JOL they had given each to item on Trial 1. Participants were given both written and verbal instructions about the T_1J. The instructions were: "Now we want you to try to remember the JOL you gave each item on Trial 1. This time when you make your judgment enter in the same number that you gave that item on Trial 1." They were asked to write down what they thought they should be making the judgment about. Three judges unanimously agreed that 9 participants did not respond appropriately, and hence were excluded from the analyses leaving only participants who clearly understood what they were doing. Because we wanted to be sure they did not rely on information from the Trial 1 test, participants were only tested on the second trial.

Results

Planned comparisons showed that recall performance for the 1–5 ($M = .49$, $SE = .10$) and 5–1 ($M = .57$, $SE = .06$) conditions were not significantly different from one another, $t(23) = 1.84$, $p > .05$, as in previous experiments.

Was memory for Trial 1 JOLs (T_1J) accurate?

Planned comparisons with the 1–5 and 5–1 conditions for the T_1J judgments showed that the 1–5 condition ($M = .32$, $SE = .04$) was significantly lower than the

5–1 condition ($M = .49$, $SE = .10$), for a difference of .17, $t(23) = 4.79$, $p < .001$, $CI_{.95} = .09, .24$. These results indicated that people were able to remember that they had given lower JOLs to items given only 1 repetition on Trial 1 than to items given 5 repetitions. They provide some support for the idea that people might have based Trial 2 JOLs on remembered Trial 1 JOLs.

The gamma correlation between the T_1J and the Trial 1 JOLs was relatively high ($M = .59$, $SE = .05$). However, the T_1J gamma was significantly lower than the gamma assessing MPT judgment accuracy in Experiment 3 ($M = .99$, $t(45) = 8.12$, $p < .001$). People's memory of their prior test performance showed superior accuracy to their memory for their prior JOLs.

Would use of T_1J result in Trial 2 underconfidence?

The 1–5 ($M = -.17$, $SE = .06$, $t(23) = 2.57$, $p = .02$) and 5–1 ($M = -.09$, $SE = .04$, $t(23) = 2.16$, $p = .04$) T_1J judgments were both significantly underconfident, so the general answer is 'yes.' However, the T_1J judgments were not significantly more underconfident for the 1–5 condition than the 5–1 condition, ($t < 1$, $p > .05$), contrasting with a pattern found consistently across Experiments 1 and 2. This comparison suggested that memory for Trial 1 JOLs may not have been the source of the UWP effect.

Would T_1J lead to improvements in resolution by Trial 2?

The mean gamma correlation between Trial 2 JOL and Trial 2 test in Experiments 1, 2 and 3 ($M = .62$, $SE = .03$) was significantly higher than the mean gamma between T_1J and Trial 2 test in Experiment 5a ($M = .41$, $SE = .09$), for a difference of .21, $t(108) = 2.67$, $p = .01$, $CI_{.95} = .05, .36$. This comparison indicated that, while people were able to remember their Trial 1 JOLs, that information alone was not adequate to generate gamma correlations similar in magnitude those found between Trial 2 JOLs and Trial 2 test as in experiments that demonstrate the UWP effect.

Discussion

Experiment 5a showed that even when total repetitions and, consequently, recall performance was matched people were generally able to remember that they had given lower JOLs to items that had received one repetition on Trial 1 in comparison to items that had received 5 repetitions. Additional criteria were needed to show that this Trial 1 difference could produce the UWP effect. We evaluated whether memory for Trial 1 JOLs could account for the same calibration patterns and gamma correlations found with Trial 2 JOLs in our earlier experiments. We found that it could to some extent but not entirely. While underconfidence arose across conditions, a comparison of the calibration levels for the 1–5 and 5–1 conditions revealed no differences, whereas across Experiments 1 and 2, the condition with fewer Trial 1 repetitions (or time) consistently showed greater underconfidence. Finally, gamma correlations for judgments based on Trial 1 JOLs were substantially lower than gamma correlations for JOLs that followed a test. Use of Trial 1 JOLs would not result in the increased resolution seen in the UWP effect. Despite the availability of prior JOLs as cues, the data suggest that people may not be using this information to inform their Trial 2 JOLs.

Experiment 5b

Experiment 5a indicated that people could remember their Trial 1 JOLs. Was this the source of the UWP effect? To further address this question, in Experiment 5b we eliminated the first trial JOLs, but administered a Trial 1 test. While it was possible that people made JOLs even though they had not been asked to, our goal was to assess whether UWP occurred in the absence of explicitly made Trial 1 JOLs, and when there was a Trial 1 test. If Trial 1 JOLs were responsible for the UWP effect, then without a Trial 1 JOL, the Trial 2 JOLs should not be biased. If past test performance was sufficient to influence the Trial 2 judgment, then Trial 2 JOLs should demonstrate both a downward calibration bias and improved resolution, that is, they should show the classic pattern.

Method

The participants were 24 undergraduates at Columbia University and Barnard College. The design, materials and procedure were identical to those in Experiment 5a except for the following: People did not make Trial 1 JOLs. They were however given a Trial 1 test. Instructions for making the Trial 2 JOL came at the start of the second trial. JOLs were explained as in Experiment 1.

Results

A planned comparison on the 1–5 and 5–1 conditions showed a large difference in Trial 1 recall performance between the 1–5 ($M = .14$, $SE = .03$) and the 5–1 group ($M = .58$, $SE = .04$), for a difference of .44, ($t(23) = 9.61$, $p < .001$, $CI_{.95} = .34, .53$), but there was no difference between these two groups on Trial 2 performance, ($t < 1$, $p > .05$).

Trial 2 JOLs were significantly lower for the 1–5 condition ($M = .60$, $SE = .05$) than for the 5–1 ($M = .69$, $SE = .04$) condition, by a difference of .09, $t(23) = 3.19$, $p < .01$, $CI_{.95} = .03, .14$, despite matched Trial 2 performance, and in the absence of Trial 1 JOLs, offering another demonstration that MPT is used in making Trial 2 JOLs.

As in the earlier experiments, the 1–5 and 5–1 conditions showed a significant .083 difference in Trial 2 underconfidence, $t(23) = 2.80$, $p = .01$, $CI_{.95} = .02, .14$. The 1–5 condition showed significantly more underconfidence ($M = -.08$, $SE = .04$, $t(23) = 1.87$, $p = .04$) than the 5–1 condition ($M = .003$, $SE = .04$), which was not significantly different from zero ($t < 1$, $p > .05$).

The gamma correlations between Trial 2 JOLs and the test on Trial 2 were not different for the 1–5 ($M = .55$, $SE = .07$) and 5–1 conditions ($M = .70$, $SE = .06$), $t(18) = 1.68$, $p > .05$. A comparison of the magnitude of these Trial 2 gamma correlations with those found in the earlier experiments revealed no significant differences, ($t(107) = 1.01$, $p > .05$) suggesting that memory for the Trial 1 test alone could produce the high Trial 2 gamma correlations.

Discussion

In the absence of the Trial 1 JOLs, but in the presence of a Trial 1 test, the Trial 2 JOLs were both underconfident and showed high resolution to test. These results were like those in Experiments 1 and 2 which both had a Trial 1 JOL and a test phase. Trial 1 test alone was sufficient to produce the UWP effect.

General discussion

These experiments implicate memory for past test as a heuristic for making immediate judgments of learning after the first trial. They also showed that this heuristic provides an account of the UWP effect that encompasses both the negative calibration biases in combination with improved resolution. In Experiments 1 and 2 we manipulated test performance on the first trial, while equating it on the second trial. It was shown with repetitions in Experiment 1, that second trial JOLs were influenced by first trial test performance, positively implicating the MPT heuristic. When we manipulated Trial 1 test performance with the amount of study time items were given in Experiment 2, we observed the same pattern, as predicted by a reliance on MPT. We confirmed that items that were forgotten on the past test but remembered on the next test, contribute disproportionately to the underconfidence effect.

The question remained, at the end of Experiment 1 and 2, as to whether this reliance on past test performance really was used in making the judgments, or whether some other correlate of Trial 1 was responsible. Experiments 3, 4 and 5 contrasted three potential sources of Trial 1 bias. This allowed us to assess the potential contribution of (1) memory for past test, (2) memory for Trial 1 encoding fluency, and (3) memory for Trial 1 JOLs to the UWP effect. In Experiment 3 we showed that people do have extraordinarily good memory for their performance on the prior test. If they were to use this alone, they would show a pattern of underconfidence accompanied by improvements in gamma correlations, like that exhibited in the UWP effect.

Surprisingly, the predictive accuracy of performance on the upcoming trial was better with the MPT judgments about the prior trial than when people made JOLs about what they would do on the upcoming trial. One reason for the disparity may have been the different scales that were used to make MPT judgments and JOLs. MPT judgments were binomial while JOLs ranged from 0 to 100. Previous research (Dunlosky & Nelson, 1994, 1997; Weaver & Kelemen, 1997) has shown that people use intermediate values, rather than only the highest and lowest values, when making immediate JOLs (though they rely more exclusively on extreme values when making delayed JOLs). Consistent with past reports, our participants used the intermediate values in the JOL condition in Experiment 3, indicating that people in the JOL conditions were, at least some of the time, doing something more than simply deciding whether they recalled or did not recall the item on the last test when making the JOLs. Analyses in Experiments 1 and 2 signaled that inadequate assessments of new learning might play a role. It therefore seems likely that people use their knowledge of past test performance, but that additional information, perhaps a muted appraisal of new learning, is also implicated in the judgment.

In Experiments 4 and 5a we attempted to rule out a number of potential explanations of the UWP effect. Experiment 4 ruled out memory for Trial 1 encoding fluency differences. People were not able to distinguish Trial 1 encoding differences as evidenced by identical encoding fluency judgments for items in the 1–5 and 5–1 conditions. In Experiment 5a we tested whether people had access to the JOLs they had given each item on Trial 1. People were generally able to make accurate judgments about their Trial 1 JOLs, however, calibration levels for the 1–5 and 5–1 conditions were not different, and gamma correlations between the T_1J judgments and recall performance were lower than the correlations we found between Trial 2 JOLs and Trial 2 test in the earlier experiments. Accurate memory for the JOLs in the absence of the appropriate calibration and resolution magnitudes indicated that people could remember their past trial JOLs, but it was not sufficient to account for the UWP effect. Furthermore, in Experiment 5b, we showed that all aspects of the UWP effect were in evidence when no Trial 1 JOL was made and people therefore could not possibly use that information. As long as there was a Trial 1 test, Trial 2 JOLs showed a difference between the Trial 2 JOLs for the 1–5 and 5–1 conditions, underconfidence and high resolution.

While these data point to the MPT heuristic, there are several other possible explanations that should be considered. Scheck and Nelson (2005) proposed that the UWP effect might reflect an anchoring and adjustment of judgments toward a psychological anchor determined by the overall level of recall. They proposed that UWP obtains when recall is higher than the anchor, which exerts a downward pull on the JOLs. We can try to apply a modified anchoring and adjustment explanation to our basic effect in Experiments 1 and 2 if the anchor were determined not by the current, Trial 2, level of recall, which was the same in the two treatment combinations of interest (so their original explanation–which presumably relies on second trial recall levels—cannot account for these data) but by the differential performance on Test 1. This explanation might then propose that the condition with low Test 1 performance (1–5 in the first experiment or 1–8 in the second) produced lower Trial 2 JOLs than did the conditions with higher Test 1 performance, because of a lower anchor point. The lower anchor for items in the 1–5 or 1–8 conditions would be determined by overall performance in those conditions on the first test.

This potential explanation is subtly different from the MPT heuristic. While both explanations would implicate Test 1 performance, the modified anchoring explanation would do so via a general condition-wide anchor, while the MPT heuristic does so on an item by item basis: people give high JOLs to items that were recalled rather than not recalled on the past test. There just happen to be fewer of these previously recalled items in the 1–5 than the 5–1 condition, which is why the overall condition shows the mean effect on JOLs. The modified anchoring view would say that people can remember which condition a particular item came from and modulate their JOL based on that condition's anchor. There are no data indicating whether or not people are able to remember correctly from which treatment combinations items came, though we do know that source judgments of this sort are difficult. We do know, given the results of Experiment 3 and a number of past experiments (e.g., Gardiner & Klee, 1976), that people can remember very well whether they were right or wrong on particular items in a past test, as is necessary for the MPT hypothesis.

A second implication of the anchoring hypothesis is that the anchor should exert a fairly uniform pull on all items in the conditions on which the anchor is based. In contrast, the MPT heuristic says that items that were forgotten and remembered on Trial 1 test should show quite different Trial 2 JOLs (regardless of treatment combination). The very high backward gamma correlations (which were higher than both the correlations between forward gammas on Trial 1 and 2), suggest that people used memory for particular items, not two global anchors. These item-based correlations count against the anchoring view, as it has been modified and articulated to explain our basic data, instead, favoring the MPT heuristic. Similarly, Finn and Metcalfe's (2007) findings, as well as those from the same analysis reported here, show that the FR items, those that were not recalled on the Trial 1 test, but were recalled on the Trial 2 test, show a selective and disproportionate underconfidence effect, across repetition conditions. These data also go against the idea that the anchors are uniformly pulling down the confidence of all of the items in the conditions to which they apply.

Past test performance is an available and diagnostic source of information for second trial JOLs, as a number of researchers have argued (Gardiner & Klee, 1976; King et al., 1980; Lovelace, 1984; Mazzoni & Cornoldi, 1993). Like them, we found improvements in gamma correlations from Trial 1 to Trial 2 when there was an influence of prior test. Koriat (1997), Koriat et al. (2002, 2006) have argued that this improved accuracy may be indicative of a modification in cue use toward cues that are sensitive to the individual's particular learning experience. We concur with the general conclusion of Koriat and his colleagues that the kinds of cues that people use over successive trials become more diagnostic. How else could resolution improve? But our data indicate that one highly salient and important cue that people use is whether or not they got an item right on the last test, and that this cue, besides being diagnostic, may also be responsible, at least in part, for the underconfidence with practice effect seen on Trials 2 and beyond.

Is the MPT heuristic the only factor responsible for underconfidence? Of course it is unlikely that the MPT heuristic alone would account for all demonstrations of undercondfidence. Underconfidence can certainly arise in situations outside the paradigm we focus on here. For example, people tend to be more underconfident about their visual perception skills than their cognitive skills (Baranski & Petrusic, 1995), women tend to be less confident in their cognitive abilities than men (Rammstedt & Rammsayer, 2002), people with obsessive compulsive disorder are more underconfident than normals and other patient populations across a variety of tasks (Dar, Rish, Hermesh, Taub, & Fux, 2000), experts sometimes demonstrate more underconfidence than amateurs (Önkal, Yates, Simga-Mugan, & Öztin, 2003) and climbers on Mt. Everest are less confident in their prospective memory after having reached extreme altitudes (Nelson et al., 1990). Thus, we do not wish to claim that memory for past test is the only reason that underconfidence might be found.

In the domain of learning, underconfidence can sometimes be found when items are simply repeated without an intervening test. But the accompanying resolution improvements that distinguish the UWP effect are not present (Koriat, 1997; Meeter & Nelson, 2003, but

see Hertzog et al., 2002). If people are not tested then logically they could not rely on their memory for past test to make their JOLs, and consequently they could not use this predictive cue to improve their gamma correlations. During the course of our investigation of the basis of the UWP effect, we conducted an experiment in which participants made JOLs after each of two study trials without an intervening test. As in other studies, we found underconfidence in the repeated versus the unrepeated condition. Indeed, in our experiment, JOLs did not change from the first JOL to the second JOL when there was no intervening test. Kelley (personal communication, December 10, 2006) and Meeter and Nelson (2003), using similar paradigms, also showed that JOLs are essentially flat when there is no test between study trials. This finding makes sense: people lacked adequate information on which to base the first JOL, but without a test they also lacked information for the second. Thus, both judgments were the same. That people did not take into account the memorial effect of repeated trials in making their JOLs is consistent with much data presented by Koriat (1997) and Koriat, Bjork, Sheffer, and Bar (2004) indicating that people are often impervious to factors that influence memory such as study time and retention interval in making these judgments. These no-test JOL results demonstrate compellingly, even in a multi-trial paradigm, that reliance on the MPT heuristic is not the only reason why a person might be underconfident. But we do suggest that without a test, resolution improvements are also not likely to occur.

Finally, while underconfidence is usually taken to be maladaptive, when it is combined, as it appears to be here and elsewhere, with an increase in resolution, or predictive accuracy about which items will or will not be recallable, it may serve an important function in learning. This 'mismeasure' may help people in refocusing attention to items—those that they could not remember on the last test—that could benefit from an additional study. Items that were answered incorrectly on Trial 1 test but correctly on Trial 2 may well be items in a tenuously learned state or in what Metcalfe (2002) and Metcalfe and Kornell (2003, 2005) termed the Region of Proximal Learning. That additional study practice could benefit just these items is a possibility that we plan to investigate.

Previous research has suggested that underconfidence associated with immediate JOLs may stem from multiple sources (Koriat et al., 2002, 2006). The experiments discussed here indicate the MPT heuristic can account for the patterns of underconfidence, increase in resolution, and the higher backward correlations with past test than with the upcoming test shown with immediate JOLs. The MPT heuristic suggests that metacognitive assessments are updated over the course of learning, and specifically, that immediate JOLs are modified by past test experience.

It seems quite plausible that students study and self-test themselves several times before their final exam. Thus, multiple trial experiments may best reflect the stages that underlie the most common learning scenarios. Our experiments attempt to unravel the components involved in metacognitive assessments about how learning is proceeding and may be of substantial importance for whether that learning will continue.

References

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology, 49*, 397–407.

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.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit Memory and Metacognition* (pp. 309–338). Mahwah, NJ: Erlbaum.

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

Dar, R., Rish, S., Hermesh, H., Taub, M., & Fux, M. (2000). Realism of confidence in obsessive-compulsive checkers. *Journal of Abnormal Psychology, 4*, 673–678.

Dunlosky, J., & Serra, M. J. (2006). *Is the influence of test trials on judgments of learning analytic?* Poster presented at the 47th Annual Meeting of the Psychonomic Society, Houston, TX.

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.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about strategy effectiveness: a componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*, 462–474.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of JOLs to various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545–565.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for Judgments of Learning (JOLs) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language, 36*, 34–49.

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.

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. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22–34.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17*, 209–225.

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.

Kintsch, W. (1970). *Learning, memory, and conceptual processes*. New York: John Wiley and Sons.

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

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, Cognition, 31*, 187–194.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory and Cognition, 34*, 959–972.

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

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478–492.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69.

Koriat, A., 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.

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

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122*, 47–60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*, 196–204.

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

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

Metcalfe, J. (1998). Cognitive optimism: self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2*, 100–110.

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

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

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

Murdock, B. B. (1974). *Human memory: Theory and ata*. LEA: Mahwah.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling of knowing predictions. *Psychological Bulletin, 95*, 109–133.

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.

Nelson, T. O., Dunlosky, J., White, D. M., Steinberg, J., Townes, B. D., & Anderson, D. (1990). Cognition and metacognition at extreme altitudes on Mount Everest. *Journal of Experimental Psychology, General, 119*, 367–374.

Önkal, D., Yates, F. J., Simga-Mugan & Öztin, S. (2003). Professional vs. amateur judgment accuracy: the case of foreign exchange rates. *Organizational Behavior and Human Decision Processes, 91*, 169–185.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement*, 76.

Rammstedt, B., & Rammsayer, T. H. (2002). Gender differences in measured and self-estimated trait emotional intelligence. *Sex Roles, 42*, 449–462.

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.

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

Weaver, C. A., III, & Kelemen, W. L. (1997). Judgments of learning at delays: shifts in response patterns or increased metamemory accuracy? *Psychological Science, 8*, 318–321.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41–44.