Journal of Experimental Psychology:
Learning, Memory, and Cognition
2006, Vol. 32, No. 3, 595–608

0278-7393/06/$12.00 DOI: 10.1037/0278-7393.32.3.595

Exploring a Mnemonic Debiasing Account of the Underconfidence-With-Practice Effect

Asher Koriat, Hilit Ma'ayan, Limor Sheffer
University of Haifa

Robert A. Bjork
University of California, Los Angeles

Judgments of learning (JOLs) underestimate the increase in recall that occurs with repeated study (the underconfidence-with-practice effect; UWP). The authors explore an account in terms of a foresight bias in which JOLs are inflated when the to-be-recalled target highlights aspects of the cue that are not transparent when the cue appears alone and the tendency of practice to alleviate bias by providing learners with cues pertinent to recall. In 3 experiments the UWP effect was strongest for items that induce a foresight bias, but delaying JOLs reduced the debiasing effects of practice, thereby moderating the UWP effect. This occurred when delayed JOLs were prompted by the cue alone (like during testing), not when prompted by the cue-target pair (like during study).

Keywords: judgments of learning, underconfidence, debiasing, foresight bias

Koriat, Sheffer, and Ma'ayan (2002) documented a phenomenon that they termed the underconfidence-with-practice (UWP) effect: When participants are presented with the same list of paired-associates for several study-test cycles, their judgments of learning (JOLs) exhibit relatively good calibration on the first study-test cycle, with a tendency for overconfidence. However, a shift toward marked underconfidence occurs from the second study-test cycle onward. The UWP effect was found to be very robust, surviving several experimental manipulations. It has also been replicated in subsequent experiments since (Dougherty & Barnes, 2003; Meeter & Nelson, 2003; Scheck & Nelson, 2005; Serra & Dunlosky, 2005; Simon, 2003;Tiede, Lee, & Leboe, 2004).

The UWP effect is surprising for several reasons. First, it stands at odds with the general tendency for overconfidence that has been observed in a great many calibration studies involving retrospective confidence (see Keren, 1991; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994). Second, practice, if anything, would be expected to improve calibration because after the first study-test cycle, learners have a more concrete idea about the difficulty of the task and about their recall performance than they had before. Finally, the impairment in calibration contrasts sharply with the observation that resolution actually improves steadily with practice. Whereas calibration (or absolute accuracy, see Nelson & Dunlosky, 1991) refers to the correspondence between mean JOLs and mean recall, and reflects the extent to which recall predictions are realistic, resolution (or relative accuracy) refers to the extent to which JOLs discriminate between recalled and unrecalled items. Thus, unlike the impairment in the JOL-recall (JR) calibration that occurs with practice, resolution, as indexed by the within-person JOL-recall gamma correlation, has been found to improve steadily with practice (e.g., King, Zechmeister, & Shaughnessy, 1980; Koriat, 1997; Koriat et al., 2002; Leonesio & Nelson, 1990; Lovelace, 1984; Mazzoni, Cornoldi, & Marchitelli, 1990).

Several explanations of the UWP effect have been proposed. Koriat (1997) suggested that this effect might be a manifestation of a general tendency of learners to discount the effects of extrinsic factors—factors that pertain to the conditions of learning or to the encoding operations applied by the learner. Indeed, he found JOLs to underestimate the effects of presentation duration on recall (but see Dunlosky & Matvey, 2001). The UWP effect accords with this tendency because it implies that the effects of list repetition (an extrinsic factor) are underweighted in the computation of JOLs. This account, however, does not predict the specific JR correspondence pattern observed—increased underconfidence—and does not offer a process-type explanation of this pattern.

Several additional accounts were explored by Koriat et al. (2002). One is that participants underestimate the correctness of the responses that they supplied on the preceding recall test, hence reporting lower JOLs on a subsequent presentation of the item. It was found, however, that giving participants feedback about the accuracy of their reported targets did not eliminate the UWP effect. Another account was that a fixed-rate presentation of the items might yield an underconfidence bias if learners wrongly estimate that the amount of allotted time was insufficient to memorize the items. However, the UWP effect was also observed when study time was self-paced. Finally, Koriat et al. examined the possibility that the UWP effect is another manifestation of the so called hard-easy effect—the tendency of participants to exhibit overconfidence in their answers for difficult items, but either no bias or even an underconfidence bias for easier items (e.g., Gigerenzer,

Asher Koriat, Hilit Ma'ayan, and Limor Sheffer, Department of Psychology, University of Haifa, Israel; Robert A. Bjork, Department of Psychology, University of California, Los Angeles.

This research was carried out at the Institute of Information Processing and Decision Making, University of Haifa. It was supported by a grant from the German Federal Ministry of Education and Research (BMBF) within the framework of German-Israeli Project Cooperation (DIP). We are grateful to Rinat Gil for her help in conducting the experiments.

We are indebted to John Dunlosky, Janet Metcalfe, and Sverker Sikström for their comments on a previous draft.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa 31905, Israel. E-mail: akoriat@research.haifa.ac.il

Hoffrage & Kleinbolting, 1991; Griffin & Tversky, 1992). Because actual performance improves with practice, making items "easier," the UWP effect is consistent with the hard-easy effect. Results reported by Koriat et al., however, yielded no hint that the UWP effect was any weaker for difficult items than for easy items.

Other accounts have been examined in subsequent studies. Scheck and Nelson (2005) proposed an anchoring-and-adjustment account according to which people make their JOLs by incomplete adjustment from a psychological anchor toward a target JOL value. They reasoned that if the anchor is assumed to lie somewhere between 30%–50% correct recall, then a UWP effect should be observed only when recall is above 50%. Indeed, using a Swahili–English paired associate task, they found a UWP effect for easy pairs, for which recall in the second presentation was above 50%. The difficult pairs, in contrast, did not yield such effect.

A somewhat different version of the anchoring-and-adjustment account was proposed by Simon (2003). The initial JOLs used by learners anchor subsequent estimates such that these estimates are made as insufficient increments to the initial values rather than being made as completely new assessments. Inconsistent with this account, however, Simon found that the elicitation of JOLs only in the third and fourth study cycles failed to eliminate or reduce the UWP effect.

Tiede et al. (2004) proposed that people discount the benefit of repetition particularly when there is high similarity between each exposure to study material, because they believe that repeated study is beneficial only when there is something distinctive about each exposure. Using a list-learning paradigm, however, they found that requiring learners to encode words differently across repetitions did not eliminate the UWP effect.

Serra and Dunlosky (2005) tested a retrieval fluency account of the UWP effect according to which JOLs on a second presentation of a list are based on how fluently items are recalled on the first presentation. If responses that are more slowly retrieved in one recall test elicit lower JOLs, but are actually better recalled in a subsequent test (see Benjamin, Bjork, & Schwartz, 1998), then these items could be responsible for the UWP effect. Although retrieval fluency in one recall test correlated negatively with JOLs on a subsequent study of the list, the UWP effect was observed for items with short retrieval latencies, as well as for those with long retrieval latencies on a preceding recall test.

Finally, Finn and Metcalfe (2004) proposed that immediate JOLs are based on the memory of one's item-specific performance on a preceding recall test, and the UWP effect derives from a failure to adequately take into account current-trial learning. Indeed, JOLs in one block were found to correlate more strongly with test performance on a previous block than with test performance on that block (see also Koriat, 1997). When the same item was repeated five times in block 2, JOLs were lower than when it was repeated five times in block 1, suggesting that participants relied heavily on first-block test performance in making JOLs in the second block.

Altogether, these results do not point to a single mechanism that can account for the occurrence of the UWP effect, suggesting that this effect might be multiply determined. In this article we explore an account that provides partial explanation of the UWP effect and specifies some of the conditions that may contribute to it.

According to our mnemonic debiasing account, the increased underconfidence with practice derives from a combination of two effects that have been previously demonstrated in connection with the monitoring of one's competence during learning. The first is the *foresight bias* (Koriat & Bjork, 2005; 2006). According to Koriat and Bjork, learners often experience an illusion of competence because they assess their degree of mastery of the studied material in the presence of information that they will be required to recall later. Thus, on a typical memory test, people are presented with a question and are asked to produce the answer. In contrast, in the corresponding learning condition both the question and the answer generally appear in conjunction, so that the prediction of one's future memory performance occurs in the presence of the answer. This difference between the learning and testing situations may produce an illusion of competence that derives from the failure to discount what the learner now knows.

Koriat and Bjork's studies indicated that the foresight bias does not occur across the board. Rather, JOLs are particularly inflated when an answer, presented during study, brings to the fore aspects of the question that are less likely to emerge during testing, when the question is presented alone. Using paired-associates, they distinguished between two types of cue-target associative relations that may influence JOLs and recall—a priori and a posteriori association. A priori association refers to the probability with which the cue word, when presented alone, brings to mind the target word (as reflected, e.g., in word-association norms). In contrast, a posteriori association refers to the perceived association between the cue and the target when *both* are present. This association is affected not only by the a priori association from the cue to the target, but also by the backward associations from the target to the cue. An illusion of knowing occurs when these associations are strong in comparison to the a priori associations. Thus, when the presence of the target highlights aspects of cue that are less apparent when the cue is presented alone during testing, recall predictions will be inflated.

Several results supported this conceptualization. Consider asymmetrically associated word pairs for which the association in one direction is strong, whereas the association in the opposite direction is relatively weak. For example, the likelihood of *cheddar* eliciting *cheese* in the word-association task is .92, whereas that of *cheese* eliciting *cheddar* is only .05 (Nelson, McEvoy, & Schreiber, 1998). Backward-associated pairs (e.g., *cheese-cheddar*) were expected to induce an illusion of competence because the association from the target to the cue inflates the a posteriori relatedness relative to the a priori relatedness. Indeed, when asymmetrically associated word pairs were presented for study in a forward direction (so that the strongest association was from the cue to the target), mean JOLs was practically identical to mean recall. In contrast, when the words appeared in a backward direction, JOLs and recall averaged 75.7% and 60.3%, respectively (Koriat & Bjork, 2005, Experiment 2). Thus, JOLs were perfectly calibrated for forward-associated pairs but were considerably inflated for backward-associated pairs.

Another demonstration of the foresight bias was obtained by using pairs with high a posteriori association but *zero* a priori association (Experiment 3). As expected, these purely a posteriori pairs yielded a particularly marked illusion of competence. In fact, the results of several experiments suggested that, in general, learners tend to perceive an association between words that are unrelated according to word-association norms. Thus, cue-target pairs with zero associative strength consistently produced inflated JOLs.

Altogether, these results support the hypothesis that an illusion of competence is likely when a posteriori associations are strong

relative to a priori associations. Such an asymmetrical pattern of associations is not uncommon because the presentation of the "answer" along with the "question" often produces the feeling that the answer is "natural" or even "obvious."

The foresight bias can help explain the weaker effects of practice on JOLs than on recall because of a second phenomenon—the *mnemonic debiasing effect of practice*. As noted earlier, it has been observed that the relative accuracy of JOLs improves steadily with repeated practice studying the same list of items (e.g., King et al., 1980; Koriat, 1997; Leonesio & Nelson, 1990; Lovelace, 1984; Mazzoni et al., 1990). It was proposed that study-test experience provides learners with mnemonic cues about the ease of learning and recalling each specific item, thereby enhancing the accuracy of recall predictions. Results reported by Koriat (1997) and Koriat, Ma'ayan, and Nussinson (2006) support this proposition. Furthermore, there is evidence suggesting that it is test experience rather than study experience that yields the greater benefit for JOL accuracy, possibly because test experience provides mnemonic cues regarding the success and fluency of retrieving the target (King, et al., 1980; Koriat & Bjork, 2006).

On the basis of these results, Koriat and Bjork (2006) hypothesized that study-test experience should help learners overcome the contaminating effects of inflated a posteriori associations by providing learners with diagnostic cues regarding the retrieval fluency of each item. Indeed, repeated practice was found to reduce the illusion of competence associated with the foresight bias. For example, (Koriat & Bjork, 2006; Experiment 1), backward-associated pairs yielded inflated JOLs on the first presentation of a list of paired associates (with JOLs and recall averaging 74% and 58%, respectively), whereas forward-associated pairs exhibited good JR correspondence. With repeated study of the list, the difference between the backward and forward pairs in JR correspondence disappeared and both types of pairs disclosed an underconfidence tendency of a similar magnitude. Thus, it is particularly the backward-associated pairs that yielded the strongest change in JR correspondence across presentations, a change from a strong overconfidence bias to an underconfidence bias. In this study we explored the possibility that the UWP effect derives in part from the beneficial effects of practice in alleviating the foresight bias for certain types of word pairs. Although the proposed mnemonic debiasing process cannot explain why an underconfidence occurs on the second study-test cycle of a list, it can account for one aspect of the UWP effect—the change in calibration that occurs with practice in the direction of underconfidence. This change results in JOLs increasing less steeply with practice than does recall, thus underestimating the beneficial effects of practice on recall.

In Experiment 1 we examined the hypothesis that the UWP effect derives primarily from the effects of practice in reducing the inflated JOLs associated with items with inordinately strong a posteriori associations. Three types of paired associates were used: purely a posteriori pairs, a priori pairs, and unrelated pairs. Assuming that the a posteriori pairs induce a foresight bias, and that this bias tends to be remedied by practice, we should expect that the reduction in overconfidence with practice should be particularly pronounced for these pairs.

In Experiments 2 and 3 we test the mnemonic-debiasing account of the UWP effect by introducing a second debiasing procedure that has been found to yield the same type of effects as those of practice—delaying JOLs until a few trials after study. Delayed JOLs, prompted by the stimulus alone, have been found to be markedly more accurate than immediate JOLs (e.g., Nelson & Dunlosky, 1991). Nelson, Narens, and Dunlosky (2004), as well as Koriat and Ma'ayan (2005), reported evidence suggesting that when JOLs are delayed learners base their recall predictions on the success and ease with which the to-be-remembered items are accessed, and ease of access is a more valid cue for subsequent recall when JOLs are delayed than when they are made immediately after study.

Experiment 2 is predicated on the assumption that a similar process underlies the effects of practice and delaying JOLs—the availability and use of mnemonic cues that are diagnostic of recall. Therefore, the two manipulations constitute alternative means for achieving the same end so that delaying JOLs should alleviate the foresight bias that is observed on the first presentation of a list and consequently result in a more moderate reduction in JOLs with practice. In Experiment 2 we used forward-associated, backward-associated and unrelated word pairs as in Koriat and Bjork (2005; Experiment 2). The list was presented for three study-test cycles, and JOLs were elicited either immediately after study or after some delay. We expect the strongest UWP effect to occur for the backward pairs when JOLs are solicited immediately after study. For these pairs, however, delaying JOLs is expected to reduce the foresight bias already on the first presentation of the list, thereby eliminating the reduction in overconfidence that occurs with practice. Experiment 3 examined the possibility that the elicitation of delayed JOLs in response to the cue-target pair, rather than in response to the cue alone, would not eliminate the UWP effect because it does not allow learners to experience the ease with which the target comes to mind and therefore does not help protect against the contaminating a posteriori associations emanating from the target.

Experiment 1

Method

Participants. Forty Hebrew-speaking University of Haifa undergraduates participated in the experiment: 8 were paid for participation and 32 received course credit.

Materials. The list of paired associates was the same as that used in Koriat and Bjork (2005, Experiment 3). It included 72 Hebrew word pairs, comprising 24 pairs with a high a priori association, 24 purely a posteriori pairs (to be referred to as "a posteriori" pairs), and 24 unrelated pairs. The high-association pairs were taken from Hebrew word association norms, such that the target word was a common response to the cue word. (The average probability of association across the 24 pairs was .21). The a posteriori pairs were selected by two judges to be semantically or associatively related, but their a priori association according to the norms was zero. Examples (translated from Hebrew) are: *clean-soap*, *bed-night*, *laugh-humor*. Finally, the unrelated pairs had zero association or were judged as unrelated.

Apparatus. The experiment was conducted on a personal computer. The stimuli were displayed on the computer screen. JOLs and recall, spoken orally by the participant, were entered by the experimenter on a keyboard.

Procedure. The experiment included two study-test cycles. Participants were instructed to study 72 paired-associates and to assess the chances that they would be able to recall the target word in response to the cue word in a subsequent test that would take place immediately after the whole list had been presented. Each study trial began with a cross at the center of the screen, accompanied by a beep. The cross, which appeared for 500 ms, was replaced by a presentation of the cue-target pair for 2 s. After

the disappearance of the pair, a JOL prompt appeared at the bottom of the screen: "The chance to recall (0%-100%) _____," participants reported their estimate orally, and the experimenter entered the data on her keyboard.

Participants were given a 5 min filler task after the study phase. In the test phase that followed, the 72 cue words appeared in a random order, and participants were required to say aloud the response word within 6 s. The entire procedure was then repeated one more time. The order of presentation of the pairs was randomly determined for each participant for each study and test phase.

Results

The UWP effect across all pairs. We shall first examine the results across all pairs to allow comparison with previous studies. These results (Figure 1, panel A) indicated that whereas JOLs and recall averaged 57.0% and 38.3%, respectively, in Presentation 1, they averaged 57.0% and 59.6%, respectively, in Presentation 2. A Presentation × Measure (JOL vs. recall) ANOVA on these means yielded $F(1, 39) = 110.13$, $MSE = 41.21$, $p < .0001$, $\eta_p^2 = .74$, for the interaction. The overconfidence bias in Presentation 1 was significant, $t(39) = 8.06$, $p < .0001$, but the underconfidence bias in Presentation 2 was not, $t(39) = 1.2$ ($p = .24$).

The pattern depicted in Figure 1 (panel A) departs from the typical UWP effect because there was little evidence for an underconfidence bias in the second presentation. This deviation supports Scheck and Nelson's (2005) claim that the underconfidence effect for the second study-test cycle of a list is not as pervasive as had been claimed (see following). The results, however, are generally consistent with the UWP effect in indicating a weaker effect of presentation on JOLs than on recall, so that JOLs underestimated the effects of learning.

Comparing the UWP effect for the a priori, a posteriori, and unrelated pairs. Figure 1 (panel B) presents mean JOLs and recall for the three types of pairs separately. A 3-way ANOVA, Pair Type × Presentation × Measure (JOL vs. recall) indicated that all main effects and interactions were significant. Of particular interest are two interactions. First, the Presentation × Measure interaction was highly significant, $F(1, 39) = 109.39$, $MSE = 123.85$, $p < .0001$, $\eta_p^2 = .74$. It can be seen that each of the three types of pairs disclosed the interactive pattern observed in panel A—a weaker effect of presentation on JOLs than on recall. Indeed, this interaction was significant for each of the three pair types: $F(1, 39) = 44.00$, $MSE = 65.23$, $p < .0001$, $\eta_p^2 = .53$; $F(1, 39) = 97.17$, $MSE = 109.53$, $p < .0001$, $\eta_p^2 = .71$, and $F(1, 39) = 45.14$, $MSE = 44.58$, $p < .0001$, $\eta_p^2 = .54$, for the a priori, a posteriori and unrelated pairs, respectively. Furthermore, for all pair types, this pattern was contributed by a significant overconfidence bias in presentation 1, $t(39) = 4.21$, $p < .0001$, $t(39) = 8.29$, $p < .0001$, and $t(39) = 6.76$, $p < .0001$, for the a priori, a posteriori, and unrelated pairs, respectively.

Second, however, these results were qualified by a triple interaction, $F(2, 78) = 20.71$, $MSE = 47.75$, $p < .0001$, $\eta_p^2 = .35$, suggesting that the change in JR correspondence with presentation differed for different pairs. To evaluate these differences, we focused on the shift in over/underconfidence across presentations. Following Serra and Dunlosky (2005; see also Finn & Metcalfe, 2004) we defined a shift score as (JOL2-Recall2) – (JOL1-Recall1), where JOL2 and Recall2 refer, respectively, to mean JOL and mean recall in Presentation 2, and JOL1 and Recall1 refer, respectively, to mean JOL and mean recall in Presentation 1. This score, which reflects the extent to which participants underestimated the effects of practice (with a negative shift disclosing a change toward underconfidence), amounted to −32.6% on average for the a posteriori pairs, −16.9% for the a priori pairs, and −14.2% for the unrelated pairs. These shift scores were all significantly different from zero, $t(39) = 9.86$, $p < .0001$, $t(39) = 6.63$, $p < .0001$ and $t(39) = 6.72$, $p < .0001$, respectively. Scheffé post hoc comparisons, however, indicated that the shift score was significantly ($p < .05$) larger for the a posteriori pairs than for the a priori and unrelated pairs, but the difference between the latter two types of pairs was not significant.

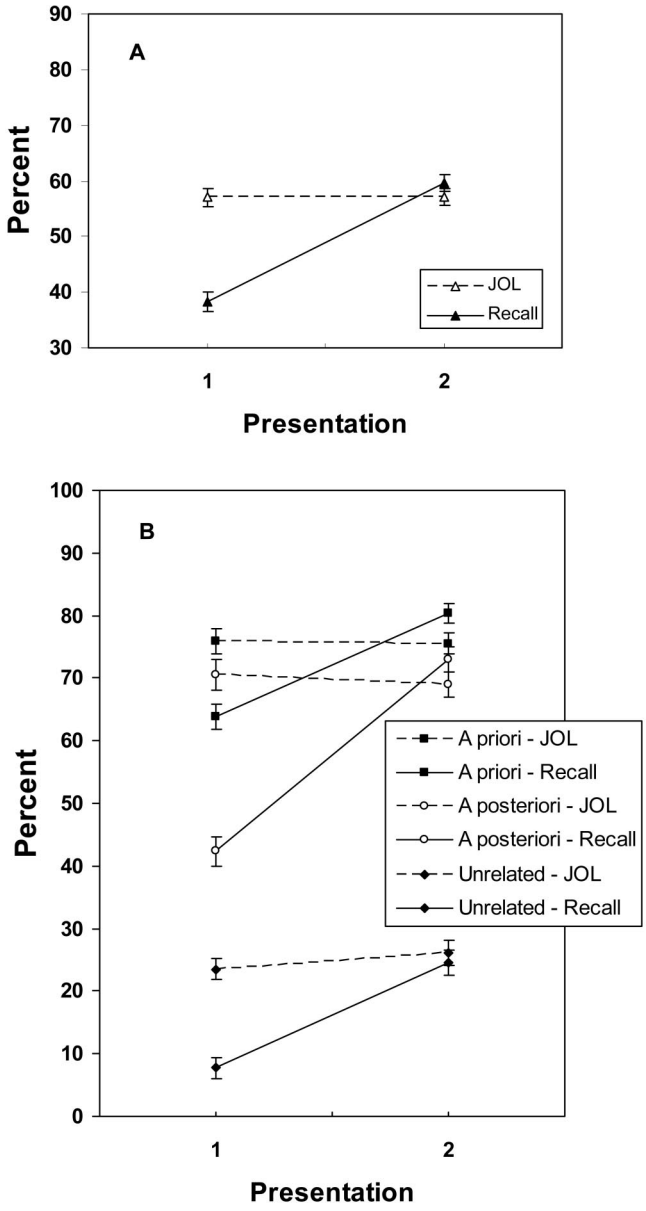

Figure 1. Mean judgments of learning (JOLs) and recall as a function of presentation, plotted across all pairs (panel A) and separately for each pair type (panel B). The error bars represent ± one standard error of the difference for within-participant comparison (see Masson & Loftus, 2003). (Experiment 1).

The comparison between the a priori and a posteriori pairs indicates that the difference between them derived almost entirely from JOLs in Presentation 1 being considerably more inflated for the a posteriori than for the a priori pairs: For that presentation, the overconfidence bias amounted to 28.2% and 12.1%, respectively for the two types of pairs. A Pair Type (a priori vs. a posteriori) × Measure (JOL vs. recall) ANOVA for that presentation yielded $F(1, 39) = 38.51$, $MSE = 67.20$, $p < .0001$, $\eta_p^2 = .50$, for the interaction. In contrast, a similar ANOVA for Presentation 2 yielded $F < 1$, for the interaction. Thus, practice helped mend the overconfidence bias to the extent that the two types of pairs exhibited about the same degree of underconfidence in Presentation 2 (4.4% and 4.9% for the a posteriori and a priori pairs, respectively), which approached significance: $F(1, 39) = 3.96$, $MSE = 218.08$, $p = .06$, $\eta_p^2 = .09$.

Discussion

The results of Experiment 1 are consistent with the UWP effect in that learners' JOLs underestimated the effects of practice. However, the JR correspondence pattern departed from the typical UWP effect (e.g., Finn & Metcalfe, 2004; Koriat et al., 2002; Serra & Dunlosky, 2005; Simon, 2003) in which an underconfidence bias is observed in Presentation 2. Only for the a priori and a posteriori pairs was there a trend toward underconfidence in that presentation. These results, as well as the differences obtained between different pair types, suggest that the UWP effect may be sensitive to characteristics of the study materials. In particular, the a posteriori pairs demonstrated the strongest decline in overconfidence with practice, consistent with the proposition that the foresight bias induced by these pairs is alleviated by practice.

Why did the results of Experiment 1 fail to reveal the full shift from overconfidence to underconfidence with practice? Several previous studies have indicated that some manipulations can enhance recall without enhancing JOLs or enhance JOLs without enhancing recall. Such manipulations either reduced or increased underconfidence overall, but curiously, the pattern of a weaker effect of practice on JOLs than on recall, as was observed in Experiment 1, was preserved. For example, increased incentives for recalling the studied items enhanced JOLs across all presentations without enhancing recall (Koriat et al., 2002). A similar pattern of a selective enhancement of JOLs was observed when participants were given feedback about the correctness of their recalls (Koriat, 1997). Also, instructing participants that repetition of a list is beneficial increased JOLs without enhancing recall (Tiede et al., 2004). In contrast, Tiede et al., (2004) found that encoding a word differently on different presentations or requiring an increasing amount of effort during study enhanced recall without affecting JOLs. All of these manipulations, however, did not modify the interactive pattern in which JOLs were less sensitive to repetition than was recall performance. Thus, perhaps some of the specific characteristics of the list of stimuli used in Experiment 1 resulted in a selective enhancement of JOLs, so that the effect of practice is expressed as reduced overconfidence rather than as increased underconfidence. We shall address this possibility after the results of Experiments 2 and 3 have been examined.

Experiment 2

Experiment 2 introduced a second debiasing procedure that has been found also to improve JOL accuracy—delaying JOLs. Nelson and Dunlosky (1991; see also Dunlosky & Nelson, 1992, 1994) found JOLs, when prompted by the cue alone, to be considerably more accurate when they were delayed until shortly after study than when they were made immediately after study. When JOLs are delayed and prompted by the cue alone, learners have the opportunity to experience the attempt to retrieve the target and can use the mnemonic cues associated with that experience as a basis for JOLs. Such cues are more diagnostic than those available when making immediate JOLs (Koriat & Ma'ayan, 2005; Nelson et al., 2004) and, in particular, are less likely to be contaminated by the inflated associations that are activated by the to-be-learned target (Koriat & Bjork, 2006).

Assuming that practice also helps alleviate the illusion of competence produced by a posteriori associations by providing learners with cues that are pertinent to retrieval fluency, then delaying JOLs should preempt the effects of practice on JR correspondence, improving that correspondence already on the first presentation, particularly for items with inflated a posteriori associations. Hence, we should expect either no UWP effect or a reduced effect for delayed than for immediate JOLs.

Indeed, several previous studies that explored the UWP effect for delayed JOLs would seem to support this hypothesis. Meeter and Nelson (2003) argued that their study was the first to establish a UWP effect for delayed JOLs, but their results indicated only a shift from a 6% overconfidence in the first presentation to a 1% underconfidence in a second presentation. In contrast, the earlier results of Dunlosky and Connor (1997) indicated little evidence for underconfidence after the first presentation, and, in fact, there was a trend toward overconfidence in some of these presentations. More recently, Finn and Metcalfe (2004) also reported that delayed JOLs did not exhibit a UWP effect. Serra and Dunlosky (2005), who conducted what is perhaps the most intensive study of the issue, found a reliable UWP effect for delayed JOLs in three experiments, but the magnitude of that effect was consistently smaller than that observed for immediate JOLs under similar conditions.

Experiment 2 was designed to test the hypothesis that the reduced UWP effect observed for delayed JOLs is specifically due to items with inflated a posteriori associations. It is particularly for such items that delaying JOLs should reduce the effects of practice on JR correspondence. Thus, we used a list of paired-associates consisting of three types of pairs as in Koriat and Bjork (2005, Experiment 2), forward-associated, backward-associated, and unrelated pairs. As noted earlier, backward pairs have been found to yield inflated JOLs, presumably because of the association from the target to the cue. The list was presented for three study-test cycles. For half of the items, participants made JOLs immediately after study, whereas for the remaining items JOLs were delayed until a few trials later. In both cases, JOLs were made in the presence of the cue alone. We expect a UWP effect for immediate JOLs that will be particularly strong for the backward pairs. These pairs are expected to yield a strong overconfidence in Presentation 1, but practice should help mend the foresight bias by reducing the inflated JOLs particularly for the backward pairs. Thus, in the second or third presentations, all three types of pairs should yield a similar over/underconfidence bias. In contrast, delayed JOLs are expected to yield little evidence for a UWP effect and, in general, little systematic change in calibration with practice for any of the pair types.

Method

Participants. Forty Hebrew-speaking undergraduates at the University of Haifa participated in the experiment: 27 were paid for participation and 13 received course credit.

Materials. A list of 72 word pairs with unidirectional association was compiled from Hebrew word association norms for college students (Rubinsten, Anaki, Henik, Drori, & Faran, 2005). These pairs were divided into two equal sets that were matched in terms of the strength of the forward and backward associations. The means of associative strength in the forward and backward directions were .39 and .04, respectively, for set A, and .39 and .05, respectively, for set B. One set was assigned to the forward direction and the other was assigned to the backward direction, with the assignment being counterbalanced across participants. In addition, 36 unrelated pairs were included. These had zero associative strength according to the norms.

Apparatus and procedure. The apparatus and procedure were the same as in Experiment 1, with the following exceptions. There were three study-test cycles. The 108 pairs were ordered randomly for each participant with the restriction that each block of 36 successive pairs included 12 forward pairs, 12 backward pairs, and 12 unrelated pairs. Of these, 6 forward pairs, 6 backward pairs, and 6 unrelated pairs were assigned to the immediate-JOL condition, and the remaining 6 pairs of each type were assigned to the delayed-JOL condition.

Participants were informed that the list included 108 paired-associates and were also instructed about the procedural difference between immediate- and delayed-JOL trials. The presentation was such that for each block of 36 pairs, for 18 randomly chosen pairs, only the cue word was shown after the disappearance of the pair, together with a JOL prompt: "The chance to recall (0%-100%) _____." In contrast, the remaining 18 delayed-JOL pairs were simply followed by the next trial, and the JOL prompt appeared only after all 36 pairs in a block had been presented—the cue word was shown together with the JOL prompt. The order of JOL elicitation for these latter pairs was such that the cue word for the first 6 pairs studied (in a block of 36) appeared first, in random order, then those of the next 6 pairs, and so on.

In the test phase, the 108 cue words appeared in a random order, and participants had to say aloud the response word within 6 s. The full study-test cycle was repeated two more times. The order of presentation of the pairs was randomly determined for each participant for each study and test phase (with the restrictions detailed above). However, the assignment of an item to the immediate or delayed condition was preserved for each participant across all three blocks.

Results

The UWP effect for immediate JOLs. We first examine the results for immediate JOLs across all items. Unlike the results of Experiment 1, those of Experiment 2 (Figure 2, panel A) disclosed a clear UWP effect, indicating a shift from overconfidence to underconfidence: A Measure (Recall vs. JOLs) × Presentation ANOVA yielded $F(2, 78) = 69.50$, $MSE = 38.54$, $p < .0001$, $\eta_p^2 = .64$, for the interaction. Whereas a significant overconfidence bias was found for Presentation 1, $t(39) = 5.55$, $p < .0001$, $\eta_p^2 = .44$, there was a significant underconfidence bias in Presentations 2 and 3, $t(39) = 3.86$, $p < .001$, $\eta_p^2 = .28$, and $t(39) = 2.13$, $p < .05$, $\eta_p^2 = .10$, respectively.

Immediate JOLs: The UWP effect for the forward, backward and unrelated pairs. Figure 2 (panel B) presents the results separately for the three types of pairs. A 3-way ANOVA, Pair Type (3) × Presentation (3) × Measure (2) yielded a nonsignificant effect for measure, $F(1, 39) = 1.40$, $MSE = 498.24$, ($p = .24$). All other effects, however, were significant. As in Experiment 1, the Presentation × Measure interaction was highly significant, $F(2, 78) = 69.17$, $MSE = 115.55$, $p < .0001$, $\eta_p^2 = .64$. All three pair types disclosed basically the same interactive pattern—a shift from overconfidence in Presentation 1 to underconfidence in Presentations 2 and 3. The Presentation × Measure interaction was significant for each of the three pair types: $F(2, 78) = 70.04$, $MSE = 79.14$, $p < .0001$, $\eta_p^2 = .64$, for the backward pairs; $F(2, 78) = 12.13$, $MSE = 85.31$ $p < .0001$, $\eta_p^2 = .24$, for the forward pairs; and $F(2, 78) = 38.05$, $MSE = 68.19$, $p < .0001$, $\eta_p^2 = .49$, for the unrelated pairs.

The triple interaction, however, was also significant, $F(4, 156) = 10.08$, $MSE = 58.55$, $p < .0001$, $\eta_p^2 = .21$, possibly

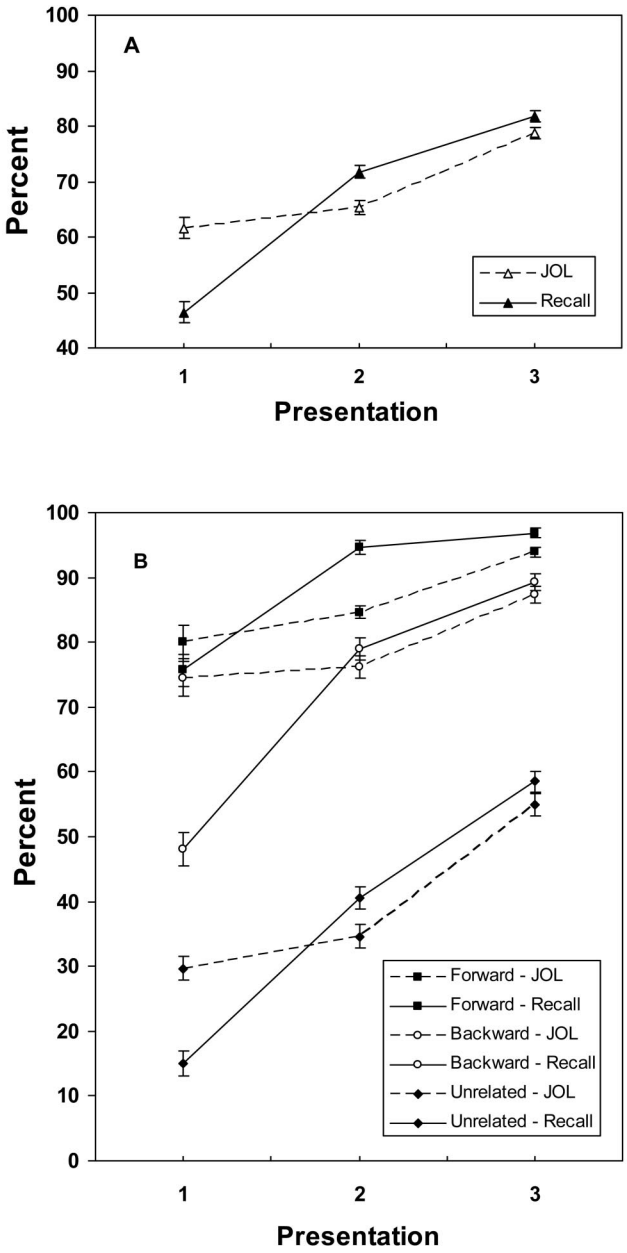

Figure 2. Mean judgments of learning (JOLs) and recall for the immediate condition as a function of presentation, plotted across all pairs (panel A) and separately for each pair type (panel B). Error bars as in Figure 1. (Experiment 2).

reflecting the observation that the shift from overconfidence to underconfidence was weakest for the forward pairs. Thus, the underconfidence shift from Presentation 1 to Presentation 2 averaged −29.3% for the backward pairs, −14.4% for the forward pairs, and −20.6% for the unrelated pairs, all significantly different from zero, $t(39) = 9.35$, $p < .0001$, $t(39) = 4.01$, $p < .001$ and $t(39) = 8.16$, $p < .0001$, respectively. Scheffé post hoc comparisons, however, indicated that the shift score was significantly ($p < .05$) larger for the backward pairs than for the forward pairs, whereas the unrelated pairs did not differ significantly from either of these two pair types. A similar pattern was obtained for the shift from Presentation 1 to Presentation 3.

Inspection of Figure 2 (panel B) suggests that the stronger underconfidence shift for the backward and unrelated pairs derived from the inflated JOLs elicited by these pairs in Presentation 1. Indeed, the overconfidence bias in Presentation 1 was significant for the backward and unrelated pairs, $t(39) = 7.09$, $p < .0001$, and $t(39) = 5.48$, $p < .0001$, respectively, but not for the forward pairs $t(39) = 1.23$, $p = .23$. Practice helped mend the overconfidence bias to the extent that the three types of pairs exhibited about the same degree of underconfidence in Presentation 3 (between 2.0% and 4.0%). For that presentation, a Measure × Pair Type ANOVA yielded $F(1, 39) = 4.49$, $MSE = 116.48$, $p < .05$, $\eta_p^2 = .10$; for measure, $F(2, 78) = 106.70$, $MSE = 321.17$, $p < .0001$, $\eta_p^2 = .73$; for pair type but not for the interaction ($F < 1$).

We should note recall performance for the backward pairs was lower than for the forward pairs, but improved more strongly from Presentation 1 to Presentation 2 than did recall performance for the forward pairs. The reason for this difference is unclear. What is important as far as calibration is concerned is the discrepancy between the effects of practice on recall and JOLs. This discrepancy was evident for all pair types, but was strongest for the backward pairs. The similarity of the JR correspondence patterns for the backward and unrelated pairs suggests that the overall level of performance does not play a critical role. This similarity, in fact, argues against Scheck and Nelson's (2005) anchoring-and-adjustment account of the UWP effect according to which this effect should be observed only when recall in Presentation 2 is above 50%. Thus, recall in Presentation 2 averaged 86.9% across the forward and backward pairs and only 40.6% for the unrelated pairs, but both types of pairs yielded a similar degree of underconfidence, with JOLs averaging 80.4% and 34.6%, respectively, for the related and unrelated pairs.

In sum, the results for immediate JOLs replicated the UWP effect, yielding a shift from overconfidence in Presentation 1 to underconfidence in Presentations 2 and 3. This shift was stronger for the backward and unrelated pairs, which had been shown to induce a foresight bias (Koriat & Bjork, 2005; 2006), than for the forward pairs. The results for these two pair types disclosed the two effects that were proposed to contribute to the UWP effect. First, a marked overconfidence bias occurred on the first study-test cycle. Second, practice helped mend this bias, possibly by providing learners with mnemonic cues that are more diagnostic of recall.

The UWP effect for delayed JOLs. We turn next to the results for delayed JOLs. Examination of the results across all items (Figure 3, panel A) indicates that now the effects of presentation are very similar for JOLs and recall. Delayed JOLs still evidenced an overconfidence bias in presentation 1, $t(39) = 6.64$, $p < .0001$, $\eta_p^2 = .53$, but this bias was apparent for Presentations 2 and 3 as well, although for these presentations it only approached significance, $t(39) = 2.01$, $p < .06$, $\eta_p^2 = .09$, and $t(39) = 1.76$, $p < .09$, $\eta_p^2 = .07$, respectively. There was a reduction in the overconfidence bias with practice, as suggested by a significant Presentation × Measure interaction, $F(2, 78) = 27.69$, $MSE = 10.35$, $p < .0001$, $\eta_p^2 = .42$.

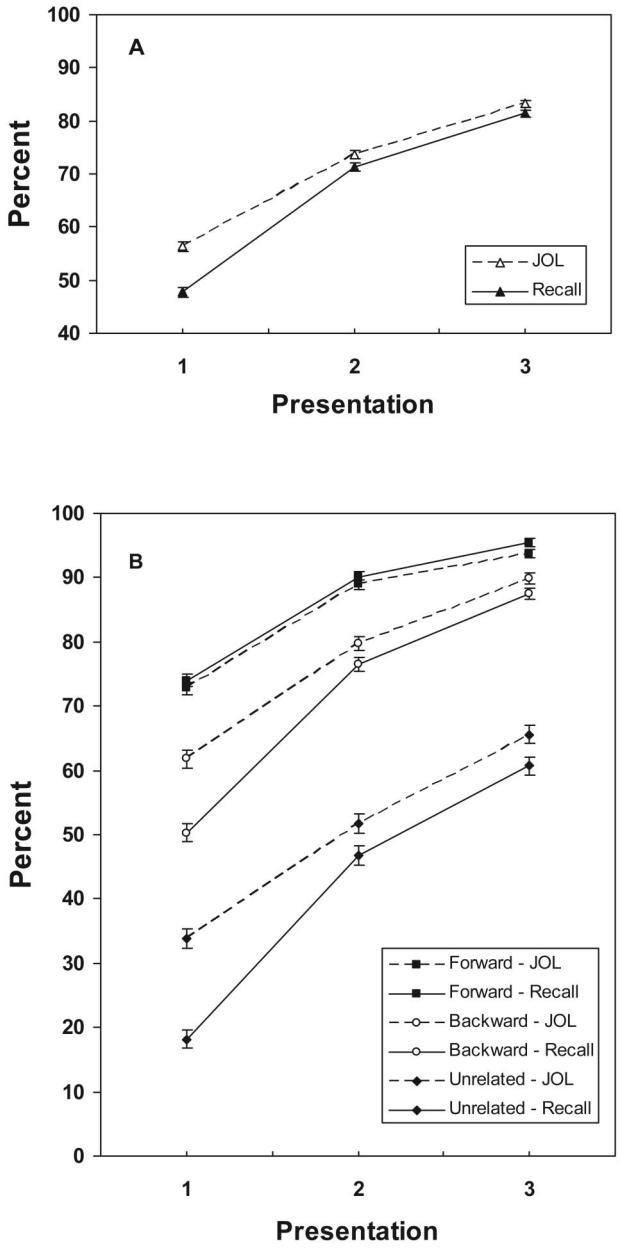

Figure 3. Mean judgments of learning (JOLs) and recall for the delay condition as a function of presentation, plotted across all pairs (panel A) and separately for each pair type (panel B). Error bars as in Figure 1. (Experiment 2).

Most important, however, in comparing the results to those of the immediate-JOL condition (Figure 2, panel A), it can be seen that the crossover interaction that was obtained for that condition (indicating a shift from overconfidence to underconfidence) disappeared. Indeed, a three-way ANOVA, Presentation × Measure × Condition (immediate vs. delayed) yielded $F(2, 78) = 33.23$,

$MSE = 18.89$, $p < .0001$, $\eta_p^2 = .46$, for the triple interaction. Thus, whereas the immediate-JOL condition yielded an underconfidence shift of −21.5% from Presentation 1 to Presentation 2, $t(39) = 10.05$, $p < .0001$, the respective shift (actually a shift toward reduced overconfidence) for the delayed-JOL condition amounted to only −6.3%, although it was still significant, $t(39) = 6.07$, $p < .0001$.

Delayed JOLs: The UWP effect for the forward, backward and unrelated pairs. Figure 3 (panel B) presents the results for the three types of pairs. Once again, A Pair Type × Presentation × Measure ANOVA indicated that all main effects and interactions were significant. The triple interaction seems to suggest the following. First, calibration was practically perfect for the forward pairs across all presentations: A Presentation × Measure ANOVA yielded a significant effect only for presentation, $F(2, 78) = 95.83$, $MSE = 102.49$, $p < .0001$, $\eta_p^2 = .71$, but not for measure, $F(1, 39) = 2.84$, $MSE = 39.89$, $p < .11$, $\eta_p^2 = .07$, or the interaction, $F < 1$. Second, the backward and unrelated pairs disclosed a very similar pattern, indicating an overconfidence bias that decreased with presentation. Thus, a Measure × Presentation × Pair type ANOVA for these pairs yielded a significant effect for Measure $F(1, 39) = 26.67$, $MSE = 227.47$, $p < .0001$, $\eta_p^2 = .41$, and for the Measure × Presentation interaction $F(2, 78) = 32.92$, $MSE = 38.23$, $p < .0001$, $\eta_p^2 = .46$. None of the interactions involving Pair type was significant (the triple interaction yielded $F < 1$). The overconfidence bias amounted to 13.6%, 4.2%, and 3.6% for Presentations 1, 2, and 3, respectively, all significant at the .05 level.

A comparison of the results for delayed JOLs with those of immediate JOLs indicates that the strongest effect of delaying JOLs was that of reducing the overconfidence bias for the backward pairs in Presentation 1. A Condition (immediate vs. delayed) × Measure × Presentation (1 vs. 2) ANOVA for the backward pairs yielded $F(1, 39) = 41.36$, $MSE = 54.01$, $p < .0001$, $\eta_p^2 = .51$, for the interaction. It is impressive, however, that delaying JOLs entirely wiped out the underconfidence bias that was found for immediate JOLs in Presentations 2 and 3.

The effects of practice and JOL delay on resolution. Let us examine the results for resolution. The assumption underlying Experiment 2 is that a similar process underlies the effects of practice and delay on JOLs—both provide participants with mnemonic cues that are more diagnostic of future recall than the cues available when immediate JOLs are solicited during the first study opportunity. Furthermore, it was assumed that each of the two manipulations can substitute for each other, so that delaying JOLs preempts the manifestation of the effects of practice and vice versa.

This idea can be tested by examining the separate and combined effects of practice and delay on resolution. Table 1 presents mean gamma correlations for immediate and delayed JOLs as a function of presentation. The results are presented separately for each pair type, as well as across all pairs. Note that gamma could not be calculated for all participants (because JOLs and/or recall averaged 100%), and therefore the number of participants on which each mean gamma was based is less than 40 in some cells. Consider first the gamma means calculated across all items. A Presentation × Condition (immediate vs. delayed) ANOVA (based only on 38 participants for whom complete data were available) yielded $F(2, 74) = 25.01$, $MSE = 0.008$, $p < .0001$, $\eta_p^2 = .40$, for the interaction. Whereas resolution improved strongly with practice for immediate JOLs, $F(2, 74) = 37.33$, $MSE = 0.013$, $p < .0001$, $\eta_p^2 = .50$, there was only a nonsystematic change for delayed JOLs, $F(2, 74) = 4.38$, $MSE = 0.008$, $p < .05$, $\eta_p^2 = .11$.

The interactive pattern observed across all items was generally obtained for each of the three pair types. Thus, it seems that the improvement in resolution that occurs as a result of delay or practice does not derive solely from increased sensitivity to differences between pair types (e.g., forward vs. backward pairs), but also from increased sensitivity to interitem differences within each class of pairs.

We should note that a similar interaction between the effects of practice (over three presentations) and the effects of delay was observed by Koriat and Shitzer-Reichert (2002) for school-age children using three study-test cycles. For cue-only delayed JOLs, the JOL-recall gamma correlations averaged .92, .94, and .88, for Presentations 1–3, respectively. In contrast, immediate JOLs, as well as delayed JOLs prompted by the cue-target pair, yielded a monotonic increase in resolution with practice, averaging .59, .80 and .85 for Presentations 1–3, respectively.

In sum, the interactive pattern between the effects of practice and the effects of delay is consistent with the idea that these two manipulations constitute alternative means for enhancing JOL

Table 1
Mean Gamma Correlations Between Judgments of Learning (JOLs) and Recall for Immediate and Delayed JOLs as a Function of Presentation

	Immediate			Delay		
	Presentation			Presentation		
Pair Type	1	2	3	1	2	3
Forward	.20 ($n = 36$)	.54 ($n = 20$)	.50 ($n = 13$)	.86 ($n = 36$)	.59 ($n = 25$)	.96 ($n = 15$)
Backward	.35 ($n = 40$)	.46 ($n = 36$)	.68 ($n = 24$)	.87 ($n = 39$)	.74 ($n = 34$)	.71 ($n = 24$)
Unrelated	.41 ($n = 30$)	.55 ($n = 40$)	.70 ($n = 37$)	.81 ($n = 30$)	.82 ($n = 39$)	.90 ($n = 40$)
All	.61 ($n = 40$)	.78 ($n = 40$)	.83 ($n = 38$)	.89 ($n = 40$)	.87 ($n = 40$)	.93 ($n = 40$)

Note. Results are presented separately for each pair type as well as across all pairs. (Experiment 2).

accuracy. We should stress, however, that because JOL accuracy for delayed JOLs was high even on the first block, the pattern of results depicted in Table 1 could also be due to a ceiling effect.

Discussion

Altogether, the results of Experiment 2 support the idea that the UWP effect derives in part from the combined operation of two processes, the foresight bias produced when the presence of the target during learning activates aspects of the cue that are less dominant when the cue is seen alone, and the mnemonic debiasing effect of practice. First, the typical UWP effect for immediate JOLs was observed across all items. Second, separate analyses of the forward and backward pairs indicated that this effect was stronger for the backward pairs. Third, delaying JOLs reduced markedly the UWP effect. In fact, it wiped out the underconfidence bias in Presentations 2 and 3 and yielded a trend toward overconfidence for the backward and unrelated pairs.

We should note, however, that although delaying JOLs reduced markedly the overconfidence bias associated with the backward and unrelated pairs in the first presentation, it did not eliminate it altogether, so that these pairs continued to evidence greater overconfidence than did the forward pairs. Also, delaying JOLs did not eliminate the reduction in confidence with practice for the backward and unrelated pairs. In contrast, for the forward pairs, delaying JOLs was sufficient to eliminate practically entirely the UWP effect, resulting in an impressively accurate calibration across all presentations.

Experiment 3

Experiment 3 was intended to provide supplementary results to those of Experiment 2 and will be reported briefly. It was similar in all respects to Experiment 2 except that JOLs were prompted by the cue-target pair rather than by the cue only. We have proposed that the solicitation of JOLs at some delay after study in response to the cue alone provides learners with a retrieval fluency cue, which is an effective predictor of subsequent recall (Koriat & Ma'ayan, 2005; Nelson et al., 2004). In contrast, when delayed JOLs are prompted by the cue-target pair, the situation is more similar to the learning situation, and, therefore, JOLs should be still susceptible to the foresight bias that is induced by the presence of the target. Consequently, when JOLs are prompted by the cue-target pair, delaying JOLs should be largely ineffective in eliminating the UWP effect. This expectation, if borne out, should strengthen the link between the UWP effect and the process that is assumed to induce a foresight bias during learning.

Method

Participants. Forty Hebrew-speaking undergraduates at the University of Haifa participated in the experiment: 23 were paid for participation and 17 received course credit.

Materials and procedure. The materials, apparatus and procedure were the same as in Experiment 2, with the exception that JOLs were prompted by the cue-target pair rather than by the cue alone.

Results and Discussion

Immediate JOLs. The results for immediate JOLs (Figure 4, panel A) yielded the typical UWP effect, as indicated by a Presentation × Measure interaction, $F(2, 78) = 51.80$, $MSE = 43.64$, $p < .0001$, $\eta_p^2 = .57$. There was a significant overconfidence bias in Presentation 1, $t(39) = 3.68$, $p < .001$, $\eta_p^2 = .26$, that changed to a marked underconfidence bias in Presentations 2 and 3, $t(39) = 4.56$, $p < .0001$, $\eta_p^2 = .35$, and $t(39) = 5.65$, $p < .0001$, $\eta_p^2 = .45$, respectively.

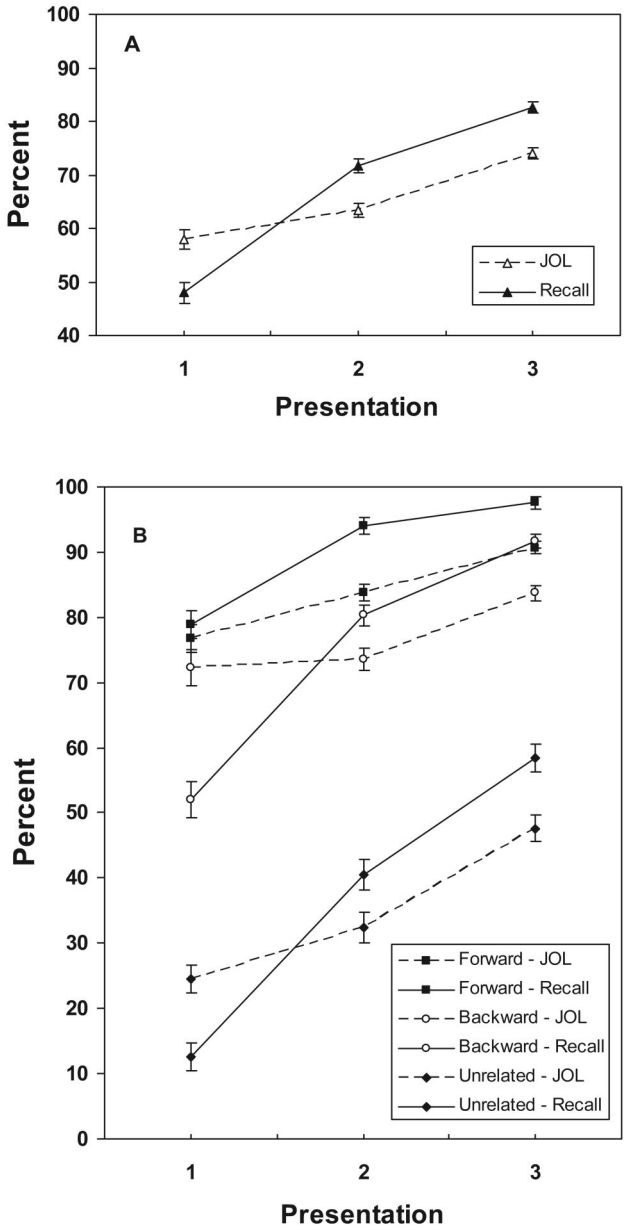

Figure 4. Mean judgments of learning (JOLs) and recall for the immediate condition as a function of presentation, plotted across all pairs (panel A) and separately for each pair type (panel B). Error bars as in Figure 1. (Experiment 3).

The results for the different types of items (Figure 4, panel B) yielded a pattern similar to that for immediate JOLs in Experiment 2, suggesting that it makes little difference whether immediate JOLs are prompted by the cue alone or by the cue-target pair. A Pair Type × Presentation × Measure ANOVA indicated that all

main effects and interactions were significant. The backward pairs exhibited a strong overconfidence bias in Presentation 1, $t(39) = 5.24$, $p < .0001$, $\eta_p^2 = .41$, that changed to underconfidence in Presentations 2 and 3, $t(39) = 2.87$, $p < .01$, $\eta_p^2 = .17$, and $t(39) = 5.16$, $p < .0001$, $\eta_p^2 = .41$, respectively. In contrast, the results for the forward pairs yielded good calibration in Presentation 1, but an underconfidence bias in Presentations 2 and 3, $t(39) = 5.67$, $p < .0001$, $\eta_p^2 = .45$, and $t(39) = 5.31$, $p < .0001$, $\eta_p^2 = .42$, respectively. The underconfidence shift from Presentation 1 to Presentation 2 was much larger for the backward pairs (−26.9%) than for the forward pairs (−8.2%), $t(39) = 8.49$, $p < .0001$, $\eta_p^2 = .65$. The respective scores for the shift from Presentation 1 to Presentation 3 were −28.2% and −4.8%, $t(39) = 9.02$, $p < .0001$, $\eta_p^2 = .68$. The unrelated pairs yielded a similar pattern to the backward pairs: An overconfidence bias in Presentation 1, $t(39) = 3.96$, $p < .001$, $\eta_p^2 = .29$, changed to underconfidence in Presentations 2 and 3, $t(39) = 2.41$, $p < .05$, $\eta_p^2 = .13$, and $t(39) = 3.64$, $p < .001$, $\eta_p^2 = .25$, respectively. The underconfidence shift from Presentation 1 to Presentation 2 (−19.9%) and from Presentation 1 to Presentation 3 (−22.7%) were both significantly different from zero, $t(39) = 7.51$, $p < .0001$ and $t(39) = 8.46$, $p < .0001$, respectively.

Delayed JOLs. Turning next to delayed JOLs (Figure 5, panel A), it appears that delaying JOLs did not modify the pattern of results substantially except that there was no overconfidence bias on the first presentation. This effect derives from an overall improvement in recall that occurred particularly in Presentation 1, possibly because delayed JOLs provided spaced retrieval practice (see Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). Thus, delaying JOL enhanced recall for each of the pair types (Figure 5, panel B). For example, a Condition (immediate vs. delayed) × Pair Type ANOVA on recall in Presentation 1 yielded a main effect of delay $F(1, 39) = 79.21$, $MSE = 69.25$, $p < .0001$, $\eta_p^2 = .67$, with recall averaging 47.8% and 57.4% for the immediate and delayed conditions, respectively. A similar ANOVA on JOLs yielded $F(1, 39) = 1.58$, $MSE = 33.11$, $p = .22$, $\eta_p^2 = .04$, for delay, with JOLs averaging 57.8% and 58.8% for the immediate and delayed conditions, respectively. The net result was a reduction in the overconfidence bias.

Apart from the improvement in recall, however, delaying JOLs did not modify substantially the overall pattern of results. Across all items, there was a significant underconfidence shift from Presentation 1 to Presentation 2, which amounted to −12.3%, $t(39) = 5.55$, $p < .0001$, although this shift was smaller than that observed for the immediate condition (−18.3%), $t(39) = 3.38$, $p < .005$, $\eta_p^2 = .23$. Also, the contrast between the forward and backward pairs was maintained: The underconfidence shift from Presentation 1 to Presentation 2 amounted to −16.0% for the backward pairs and to + 1.1% for the forward pairs, $t(39) = 6.09$, $p < .0001$, $\eta_p^2 = .49$.

Cue-only versus cue-target delayed JOLs. A comparison of Figure 3 (panel A) and Figure 5 (panel A) suggests that the cue-target presentation enhanced recall, but not JOLs, in comparison with the cue-only condition. An Experiment (2 vs. 3) × Measure × Presentation ANOVA yielded a significant Experiment × Measure interaction, $F(1, 78) = 23.75$, $MSE = 128.67$, $p < .0001$, $\eta_p^2 = .23$. Recall averaged 66.8% for Experiment 2 (cue only) and 74.3% for Experiment 3, $t(78) = 2.66$, $p < .01$. The respective means for JOLs were 71.1% and 68.5%, $t(78) = 1.10$, *ns*. It is instructive to note that whereas the cue-only delayed condition (Experiment 2) yielded a significant overconfidence bias in Presentation 2, particularly for the backward and unrelated pairs, the cue-target delayed JOLs (Experiment 3) yielded a significant underconfidence bias, $t(39) = 6.04$, $p < .0001$, $\eta_p^2 = .48$, $t(39) = 3.08$, $p < .005$, $\eta_p^2 = .20$, and $t(39) = 5.14$, $p < .0001$, $\eta_p^2 = .40$, for the forward, backward, and unrelated pairs, respectively. In sum, by and large, when JOLs were prompted by the cue-target pair rather than by the cue alone, delaying JOLs improved recall overall but did not modify the UWP pattern that was found for immediate JOLs.

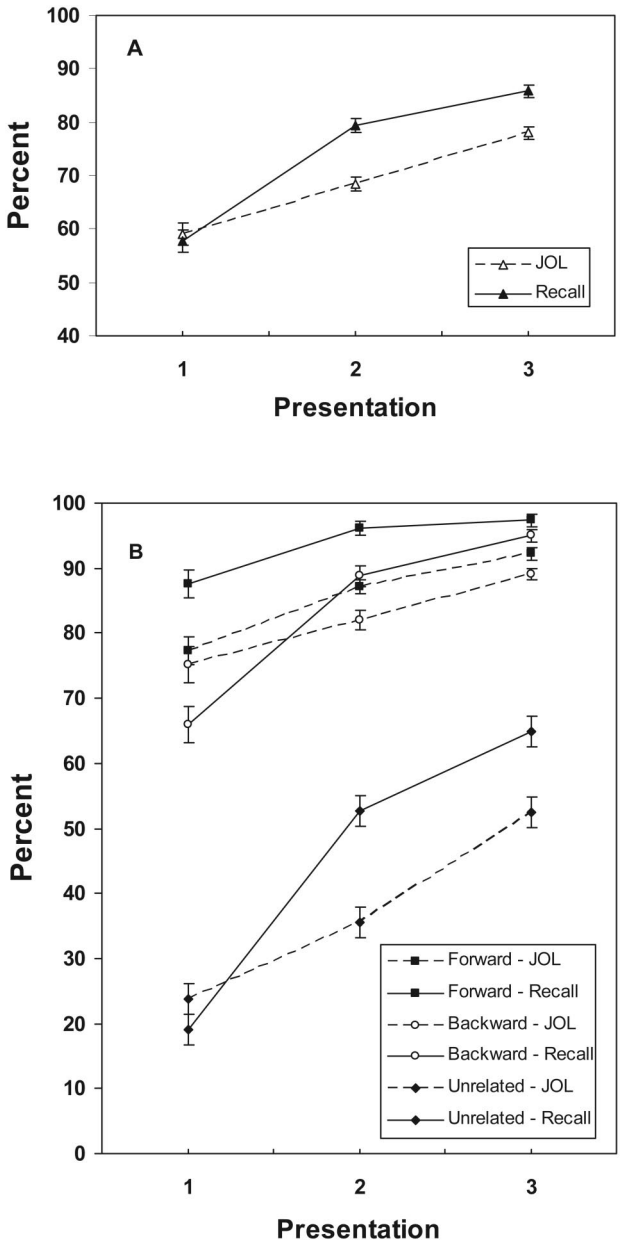

Figure 5. Mean judgments of learning (JOLs) and recall for the delay condition as a function of presentation, plotted across all pairs (panel A) and separately for each pair type (panel B). Error bars as in Figure 1. (Experiment 3).

The effects of practice and JOL delay on resolution. Practice improved JOL accuracy for both immediate and delayed JOLs. For

immediate JOLs, the JOL-recall correlation averaged .63, .79, and .86 for Presentations 1, 2, and 3, respectively. The respective correlations for delayed JOLs were .69, .80, and .86 (with n varying between 39 and 40). A Condition (immediate vs. delayed) $\times$ Presentation ANOVA (based only on 37 participants for whom complete data were available) yielded $F(2, 72) = 70.97$, $MSE = 0.013$, $p < .0001$, $\eta_p^2 = .66$ for presentation, $F(1, 36) = 2.68$, $MSE = 0.009$, $p = .12$, $\eta_p^2 = .06$, for delay, and $F < 1$ for the interaction. Thus, unlike the interactive pattern that was observed in Experiment 2 between practice and delay (see Table 1), here the effects of practice were largely similar for both types of JOLs. This pattern accords with the finding that delaying JOLs does not improve resolution markedly when JOLs are prompted by the cue-target pair (Dunlosky & Nelson, 1992) and is also consistent with the pattern observed by Koriat and Shitzer-Reichert (2002) for children.

Overall, the results of Experiment 3 support the proposition that the solicitation of cue-only delayed JOLs in Experiment 2 helped overcome the overconfidence bias for backward-associated pairs by reducing the contaminating effects of to-be-recalled target and increasing sensitivity to retrieval fluency as a cue for JOLs. When JOLs were prompted by the cue-target pair (Experiment 3), in contrast, delaying JOLs proved relatively ineffective in alleviating the foresight bias for backward-associated pairs and in eliminating the UWP effect for these pairs in comparison with immediate JOLs.

General Discussion

Overview of Hypotheses and Supportive Results

The results that have been reported so far on the UWP effect present a complex picture. On the one hand, the UWP effect seems to be quite robust. It was found for paired-associates tasks (Finn & Metcalfe, 2004; Koriat et al., 2002; Serra & Dunlosky, 2005), for a list-learning task (Koriat et al., 2002; Tiede et al., 2004) and for self-performed actions (Koriat et al., 2002). It was observed for both item-by-item JOLs as well as for aggregate JOLs, that is, estimates of the number of items that will be recalled at test (Dougherty & Barnes, 2003; Koriat et al., 2002). Furthermore, the effect survived a variety of experimental manipulations as reviewed in the introduction. On the other hand, several results seem to place constraints on the occurrence of the UWP effect. Simon (2003; see also Koriat, 1997) established that the UWP effect occurs only when the same list is repeated, not when different lists are used one after the other, suggesting that the UWP effect is specific to restudied items. Dougherty and Barnes (2003) found little evidence for a UWP effect for aggregate JOLs when attention was divided either at encoding or at retrieval. Scheck and Nelson (2005), as noted earlier, found a UWP effect only for easy paired-associates and not for difficult pairs. Even the latter pairs, however, yielded a pattern in which JOLs were less strongly affected by repeated study than was recall. Of greater relevance to the present study, our results, as well as those of others, suggest that delaying JOLs sometimes reduces or even eliminates the UWP effect (Finn & Metcalfe, 2004; Serra & Dunlosky, 2005).

By and large, however, it is the robustness of the UWP effect that makes it particularly difficult to offer a general account for this phenomenon. In the present article we proposed a mechanism that may contribute to the occurrence of the UWP effect, but that mechanism, as will be discussed later, does not provide a complete account. In this section we first review the evidence in support of the mnemonic debiasing account. In the next section we focus on observations that are not consistent with this account.

According to the mnemonic debiasing account, the UWP effect derives from a combination of two processes: (a) The foresight bias—JOLs are generally inflated because the answer, presented during study, highlights aspects of the question that are less likely to emerge during testing, when the question is presented alone (Koriat & Bjork, 2005; 2006); and (b) The debiasing effect of practice—study-test practice tends to mend the foresight bias by providing learners with mnemonic cues pertaining to retrieval fluency (e.g., Koriat & Ma'ayan, 2005). The combination of these two effects is expected to yield inflated JOLs on the first study-test block, which should be moderated by practice. The net result is that JOLs should yield an underestimation of the effects of practice on recall.

The results of Experiment 1 were generally supportive of this account. JOLs in Presentation 1 were considerably more inflated for the purely a posteriori pairs than for the a priori pairs, but the pronounced foresight bias for the a posteriori pairs was mended by practice to the extent that these pairs exhibited the same degree of underconfidence in Presentation 2 as did the a priori pairs. The unrelated pairs also yielded a pronounced foresight bias (see also Koriat & Bjork, 2005), and that bias was also alleviated by practice. The result was that JOLs increased less strongly with practice than did recall. However, the overall pattern departed from the typical UWP effect in that there was only a very slight underconfidence bias in the second study-test cycle.

The immediate-JOL condition of Experiment 2 yielded a JR correspondence pattern that was more similar to that reported by Koriat et al. (2002)—a shift from overconfidence in Presentation 1 to underconfidence in Presentations 2 and 3. Consistent with predictions, the overconfidence bias in Presentation 1 was due primarily to the backward-associated pairs, whereas the forward-associated pairs yielded good calibration (Koriat & Bjork, 2005; 2006). The backward pairs enjoyed the most benefit from practice and hence contributed a large part of the UWP effect that was observed across all items.

As expected, delaying JOLs in Experiment 2 modified the JR correspondence pattern dramatically. First, it reduced the inflated JOLs observed in Presentation 1 for the backward pairs, and second, it reduced markedly the UWP effect, particularly for the backward pairs, to the extent that there was no underconfidence bias in Presentations 2 and 3 for any of the pair types. These results are consistent with the idea that delaying JOLs, like study-test practice, helps learners overcome the contaminating effects of a posteriori associations and therefore obviate the effects of practice on the JR correspondence, particularly for the backward-associated pairs. The proposition that a similar mechanism underlies the effects of practice and of JOL delay was supported also by the interactive pattern observed for monitoring resolution (see Table 1)—practice was found to have little effect on resolution over and above that of delay.

Finally, Experiment 3 provided further support for the link between the UWP effect and the foresight debiasing effect of practice. Unlike the cue-only delayed-JOLs of Experiment 2, the cue-target delayed JOLs of Experiment 3 proved relatively ineffective in alleviating the foresight bias for backward-associated pairs and in reducing the UWP effect for these pairs in comparison

with immediate JOLs. These results support the contention that only when JOLs are solicited in response to the cue alone does delaying JOLs help overcome the contaminating associations activated by the presence of the target and moderate the magnitude of the UWP effect.

We should note that the results for the delayed-JOL condition of Experiment 2 are consistent with those of Serra and Dunlosky (2005) who found delaying JOLs to reduce, but not eliminate, the UWP effect. We also found that delaying JOLs alleviated the foresight bias for the backward-associated pairs and reduced the effects of practice on the JR correspondence for these pairs, but did not eliminate these effects entirely. This failure is, perhaps, not surprising because even when delayed JOLs are prompted by the cue alone, the target is likely to be retrievable on some occasions prior to making JOLs (see Koriat & Ma'ayan, 2005; Nelson et al., 2004), so that JOLs can still be affected sometimes by associations that are activated by the retrieved target.

Shortcomings of the Proposed Account and Some Inconsistent Results

We shall now examine some of the limitations of the proposed account. First and foremost, the mnemonic debiasing account does not explain the underconfidence observed for the second or third study-test cycles of the list. It does provide an explanation for the overconfidence bias in the first presentation and for the reduction in that bias with repeated presentations, but what is the source of the underconfidence observed after the first presentation? In fact, it is the underconfidence bias following the first presentation that has been the focus of previous proposed accounts of the UWP effect (e.g., Finn & Metcalfe, 2004; Koriat et al., 2002; Simon, 2003; Tiede et al., 2004).

Curiously enough, this theoretical problem is matched by an empirical problem. In the present study, the shift to underconfidence did not obtain consistently across all three experiments. The UWP pattern, with a significant overconfidence head and a significant underconfidence tail, was clearly replicated by the results for the immediate JOLs of Experiments 2 and 3 (Figure 2, panel A, and Figure 4, panel A). Most deviant was the pattern observed in Experiment 1 (Figure 1, panel A), in which a strong overconfidence bias was observed in Presentation 1, but the results for Presentation 2 yielded little underconfidence. How can these variations in the JR correspondence be explained?

One critical determinant seems to be the composition of the list of items used, as indicated by the variation in the JR correspondence patterns observed for different types of items (panels B in the respective figures). This variation suggests that the specific pattern obtained should depend on the nature of the stimuli used and on the composition of the entire list. In particular, our deliberate inclusion of items with inflated a posteriori associations seemed to have resulted in a more articulated overconfidence bias in Presentation 1 than has been found in Koriat et al. (2002). This comment brings to the fore the importance of representative design. Several authors have stressed the observation that in studies using general-knowledge questions the magnitude of overconfidence bias observed varies strongly with the nature of the items included in the study (Gigerenzer, et al., 1991).

In addition to the observation that the JR correspondence may differ for different types of items, we should reiterate the observation noted earlier, that several manipulations have been found to exert differential effects on JOL and recall, sometimes enhancing one without enhancing the other (e.g., Finn & Metcalfe, 2004; Koriat et al., 2002; Tiede et al., 2004). Such manipulations, of course, affect the JR correspondence. Nevertheless, the interactive pattern involving the effects of practice was observed in these studies, as well. JOLs underestimate the effects of practice on recall. In fact, a similar pattern was found in this study in comparing the results for immediate JOLs between Experiments 2 and 3. In comparison with the cue-alone presentation (Experiment 2), the cue-target presentation (Experiment 3) enhanced recall but not JOLs. This enhancement, however, did not modify the pattern of a more moderate effect of practice on JOLs than on recall.

Another example comes from aggregate judgments. When participants estimate the frequency of correct recalls across a series of paired associates, their estimates are substantially lower than the average item-by-item JOLs. As a result, aggregate JOLs yield an underconfidence bias even on the first study-test cycle. That bias, however, was also found to increase from the first study-test cycle to subsequent cycles (Koriat et al., 2002).

The variations in the JR correspondence may also affect the extent to which the effects of practice on calibration parallel its effects on resolution. In discussing the UWP effect, Koriat et al., (2002) argued that practice exerts a paradoxical effect, improving resolution while impairing calibration by instilling underconfidence. This conclusion was based on the observation that JOLs were relatively well calibrated in the first study-test cycle, but underestimated recall markedly on the second cycle. In contrast, in the present study the strongest discrepancy between JOLs and recall was sometimes observed in the first presentation, particularly for items with inordinately strong a posteriori associations. For these items, practice actually improved calibration rather than impaired it. Also in Experiment 2, delayed JOLs were found to improve calibration while also enhancing resolution. Thus, improved resolution can sometimes go hand in hand with improved calibration.

Perhaps, then, the most general description of the effects of practice on the JR correspondence is that JOLs underestimate the improvement in recall that occurs as a result of repeated study (see Serra & Dunlosky, 2005). This interactive pattern is most often instantiated in the form of a UWP pattern in which initial overconfidence gives way to subsequent underconfidence. Sometimes, however, the underconfidence segment may be missing, as in Experiment 1 of this study, or in the difficult-item condition of Scheck and Nelson (2005).

The question then remains: Although our account can explain both the initial overconfidence segment as well as the reduced confidence with practice, why is it the case that in the great majority of conditions that have been investigated, learners exhibit an underconfidence bias from the second presentation on? This is still a puzzle that invites further research, and here we may have to resort to other mechanisms that have been proposed by others (e.g., Finn & Metcalfe, 2004; Scheck & Nelson, 2005). One clue toward a solution might be found in the results for delayed JOLs in the present study. Inspection of the results presented by Serra and Dunlosky (2005) suggests that delaying JOLs not only reduces the overconfidence bias on the first presentation in comparison with the immediate-JOL condition, it also reduces the underconfidence bias in the second presentation (the latter trend is suggested by the results of their Experiments 1 and 3, Appendix A1). A similar pattern appears

in Meeter and Nelson's results (2003, Table 1). Experiment 2 of the present study clearly disclosed this pattern, too. The delayed-JOL condition of Experiment 2 yielded no sign for an underconfidence bias in Presentation 2, but instead a significant overconfidence for the backward and unrelated pairs. This pattern contrasts with the significant underconfidence bias observed for that presentation in the cue-target delayed JOLs of Experiment 3. The results of Finn and Metcalfe (2004) also indicate a trend in which immediate JOLs yield underconfidence, whereas delayed JOLs yield overconfidence in the second presentation. Thus, perhaps a detailed examination of the effects of JOL delay can shed light on the reasons for the underconfidence bias following practice.

In sum, in this study we explored a mnemonic debiasing account of the UWP effect. We capitalized on previous findings that indicated marked differences between different types of items in precipitating a foresight bias, and, indeed, the UWP effect was found to differ markedly for these item types. Furthermore, assuming that delaying JOLs also reduces the foresight bias, we showed that it also reduces markedly the UWP effect for items that are assumed to produce inflated a posteriori associations. However, although the results presented in this article are consistent with the mnemonic debiasing account of the UWP effect, that account is clearly incomplete. Possibly the UWP effect is multiply determined, requiring more than a single explanatory mechanism.

References

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

Dougherty, M., & Barnes, K. (2003). *Divided Attention and Metamemory Judgments.* Poster session presented at the 44th annual meeting of the Psychonomic Society, Vancouver, BC.

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., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1180–1191.

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

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

Finn, B., & Metcalfe, J. (2004, November). *Multitrial judgments of learning.* Poster session presented at the 45th annual meeting of the Psychonomic Society, Minneapolis, MN.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

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 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, and 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 fluency. *Memory & Cognition,* in press.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

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. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478–492.

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

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 1–17). New York: Kluwer.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 464–470.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). England: Cambridge University Press.

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

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

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

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980–94. In G. Wright (Ed.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

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

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

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., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment Recall And Monitoring (PRAM). *Psychological Methods, 9,* 53–69.

Rubinsten, O., Anaki, D., Henik, A., Drori. S., & Faran, Y., (2005). Norms for free associations in the Hebrew language. In A. Henik, O. Rubinsten, & D. Anaki (Eds.), *Word norms for the Hebrew language* (in Hebrew) (pp. 17–34). Ben Gurion University of the Negev.

Scheck, P., & Nelson, T. O. (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.

Simon, D. A. (2003). *Underconfidence with practice: Do anchoring effects play a role?* Poster session presented at the 44th annual meeting of the Psychonomic Society, Vancouver, British Columbia, Canada.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3,* 315–316.

Tiede, H. L., Lee, M., & Leboe, J. P. (November, 2004). *Investigations into the underconfidence with practice effect on judgments of learning.* Poster session presented at the 45th annual meeting of the Psychonomic Society, Minneapolis, MN.

Received July 25, 2005
Revision received November 1, 2005
Accepted December 19, 2005 ■