

## Does Retrieval Fluency Contribute to the Underconfidence-With-Practice Effect?

Michael J. Serra and John Dunlosky  
Kent State University

Judgments of learning (JOLs) made during multiple study–test trials underestimate increases in recall performance across those trials, an effect that has been dubbed the underconfidence-with-practice (UWP) effect. In 3 experiments, the authors examined the contribution of retrieval fluency to the UWP effect for immediate and delayed JOLs. The UWP effect was demonstrated with reliable underconfidence on Trial 2 occurring for both kinds of JOL. However, in contrast to a retrieval-fluency hypothesis, fine-grained analyses indicated that the reliance of JOLs on retrieval fluency contributed minimally to the UWP effect. Our discussion focuses on the status of the retrieval-fluency hypothesis for the UWP effect.

*Keywords:* metacognition, judgments of learning, underconfidence with practice, retrieval fluency, judgment bias

A judgment of learning (JOL) is a metacognitive judgment in which one predicts the likelihood of remembering an item on a criterion test (Nelson & Narens, 1990). In a typical experiment involving JOLs, a participant may study paired associates (e.g., daffodil–blood) at a fixed presentation rate, and after studying each one, he or she would predict the percent likelihood (0–100%) of recalling the second word of a pair (i.e., blood) when shown the first (i.e., daffodil). Interest in JOLs has been increasing over the past decade, partly because of their functional role in guiding self-paced study (for reviews, see Son & Metcalfe, 2000; Thiede & Dunlosky, 1999). To effectively guide self-paced study, JOLs must accurately predict criterion performance (Thiede, 1999). For instance, overconfident JOLs may lead to an inadequate duration of future study, whereas underconfident JOLs may result in excessive study. Because these kinds of bias are evident in people’s judgments in many circumstances, understanding why judgments are inaccurate has important implications for supporting effective study. Toward this end, the present research is aimed at gaining insight into why people’s JOLs become underconfident across study–test trials.

Such underconfidence of JOLs was first brought to our attention by Koriat, Sheffer, and Ma’ayan (2002). In 11 experiments demonstrating this effect (Koriat et al., 2002), participants typically studied paired associates, provided a JOL for each pair, and then were given a test of paired-associate recall. This procedure was repeated for at least two trials. Across these trials, a shift occurred

toward underconfidence in participants’ JOLs. More specifically, the mean JOL on the first trial exceeded mean recall performance, indicating overconfidence, whereas on subsequent trials criterion recall exceeded JOLs, indicating underconfidence. Koriat et al. labeled this shift the *underconfidence-with-practice* (UWP) effect because mean JOL magnitude was consistently lower than mean recall from the second trial on.

In the present experiments, we evaluated an explanation for the UWP effect first described by Koriat et al. (2002), which we refer to as the *retrieval-fluency hypothesis*. Retrieval fluency pertains to the ease with which individuals retrieve responses during a recall trial that precedes the JOLs (for a discussion of retrieval fluency as a cue for metacognitive judgments, see Benjamin & Bjork, 1996). According to the retrieval-fluency hypothesis, “the UWP reflects the failure to take into account the beneficial effects of recall experience when making JOLs” (Koriat et al., 2002, p. 160). The idea here is that participants’ JOLs are lower for responses that are retrieved more slowly than those retrieved quickly (e.g., Benjamin, Bjork, & Schwartz, 1998; Matvey, Dunlosky, & Guttentag, 2001). These lower JOLs may then underestimate subsequent recall if difficult-to-retrieve responses show a high rate of recall success on subsequent test trials. The plausibility of this hypothesis for the UWP effect is supported by results from Benjamin et al., who had participants answer general information questions (e.g., “What gem is red in color?”) during an initial trial. Immediately after answering a question, each participant also made a JOL by predicting the likelihood of recalling his or her response (e.g., “ruby”) during a subsequent test of free recall. When answering the general information questions, retrieval latencies were negatively correlated with JOLs, which demonstrated that retrieval fluency is a basis for people’s JOLs. As important, retrieval latencies were positively correlated with free recall—that is, longer retrieval times were related to a greater likelihood of free recall. Thus, retrieval fluency inappropriately biased individuals’ JOLs away from the actual level of free recall performance. Similarly, retrieval fluency may contribute to the UWP effect if retrieval latencies on Trial 1 are negatively correlated with JOLs on Trial 2, and these latencies are positively correlated with recall on Trial 2

---

Michael J. Serra and John Dunlosky, Department of Psychology, Kent State University.

Results from Experiments 1 and 2 were presented at the 44th annual meeting of the Psychonomic Society in November, 2003. Experiments 1 through 3 served as Michael J. Serra’s master’s thesis. We thank Reed Hunt and Paul Silvia for serving on the thesis committee and for their helpful comments and suggestions regarding this research. We also thank Matthew Brallier for helping to run participants.

Correspondence concerning this article should be addressed to John Dunlosky, P. O. Box 5190, Department of Psychology, Kent State University, Kent, OH 44242-0001. E-mail: [jdunlosk@kent.edu](mailto:jdunlosk@kent.edu)

(as in Benjamin et al., 1998). In this manner, retrieval fluency may inadvertently bias JOLs toward underconfidence across multiple study–test trials.

The retrieval-fluency hypothesis yields several testable predictions. First, JOLs made on Trial 2 will be negatively correlated with the speed of retrieving correct responses during Trial 1 recall. Second, the UWP effect will be largest for items in which the speed of retrieving correct responses on Trial 1 had been the slowest, and it will be smallest for items in which retrieval speed had been the fastest. This latter prediction specifically links slow retrieval with the emergence of the UWP effect. To test these predictions, we had participants study paired associates on two consecutive study–test trials and make a JOL for each item on both trials. Latencies of recalling correct responses during Trial 1 recall were recorded.

To test the first prediction, we correlated the latency of correctly retrieving responses during Trial 1 recall with JOLs on the second trial. A negative correlation was expected. It is important to note that confirming this prediction is necessary, but not sufficient, for establishing retrieval fluency as a contributor to the UWP effect. In particular, such a negative correlation can explain the UWP effect only if the fluency of retrieval during Trial 1 recall affects JOLs on Trial 2 without similarly affecting retrieval during Trial 2 recall. Accordingly, evaluating the second prediction will provide the most crucial evidence that retrieval fluency contributes to the UWP effect. To do so, we compared the underconfidence that occurred for items that were most slowly retrieved during Trial 1 recall versus those that were most quickly retrieved. The retrieval-fluency hypothesis predicts that more slowly retrieved items will show greater underconfidence.

To test these predictions, we had participants make immediate JOLs (which were originally used by Koriat et al., 2002) and delayed JOLs. For the latter, a brief delay occurred between the study and JOL for each item. Although JOL accuracy is often greater for delayed than immediate JOLs (Nelson & Dunlosky, 1991), the UWP effect has recently been demonstrated for delayed JOLs (Meeter & Nelson, 2003; Scheck & Nelson, 2005; for an exception, see Finn & Metcalfe, 2004), so perhaps retrieval fluency contributes to this effect for delayed JOLs as well. Thus, we estimated the contribution of retrieval fluency to the UWP effect for both immediate and delayed JOLs.

## Experiment 1

### Method

#### *Participants, Design, and Materials*

Sixty undergraduates from the University of North Carolina at Greensboro (UNCG) participated in this study to partially fulfill a course requirement. Participants were run 4 at a time on Apple iMac computers in the same room. These computers displayed the items to the participants and recorded all responses and latencies. The use of immediate versus delayed JOLs was a within-participant variable. Each participant studied 60 critical unrelated noun–noun word pairs (such as “daffodil—blood”). The materials, item randomization, item presentation, and procedure of the current study were identical to that of Nelson and Dunlosky (1991) with the exception that, in this instance, participants completed the procedure twice in succession.

### *Procedure*

The participants completed two trials of paired-associate study and paired-associate recall. They studied the same set of 60 pairs on both trials. For study, the computer presented each item to the participant for 6 s. For each item, participants made either an immediate or delayed JOL. The JOL prompt consisted of only the stimulus of an item (e.g., for “daffodil—blood,” the JOL prompt would be “daffodil - ?”). Participants were asked, “How confident are you that in about 10 minutes from now you will be able to recall the second word of the item when prompted with the first? (0 = definitely won’t recall, 10 = 10% sure, 20 . . . , 30 . . . , 40 . . . , and 100 = definitely will recall).” The scale responses were in increments of 10. For immediate JOLs, the JOL prompt for an item was presented immediately after the offset of the presentation of that item for study. For delayed JOLs, the JOL prompt was presented for an item at least 1 min after the offset of the presentation of that item for study, with study and JOLs for other items intervening between the study and delayed JOL for any given item (for details on list construction, see Nelson & Dunlosky, 1991).

After participants studied and judged all 60 items, they performed an unrelated distracter task for 5 min. They were then asked to recall the second word of each item (by typing it into a text field within the computer program) when prompted with the first word. Items were tested one at a time. Recall was participant paced, and omissions were accepted. The order of items was randomized anew for presentation during the recall trials. To correct for participants’ spelling and typing errors, we scored words as correct if the first three letters of a participant’s answer matched the first three letters of the correct response. As a measure of retrieval fluency, response latencies were recorded.<sup>1</sup> This procedure was repeated for a second trial in which the order of items was randomized anew for presentation at study and at test.

## *Results and Discussion*

### *The UWP Effect*

To estimate the UWP effect, we computed the signed difference between each participant’s mean JOL and mean recall performance. Mean values for both trials are reported in Table 1 along with a composite score that reflects the overall shift to underconfidence across trials (cf. Meeter & Nelson, 2003). For interested readers, we have included JOL magnitude and recall performance in Appendix A (Table A1), which are the basis of this derived measure of absolute accuracy.

Two effects are evident from inspection of Table 1 and were substantiated by a 2 (kind of JOL; immediate vs. delayed)  $\times$  2 (trial; Trial 1 vs. Trial 2) analysis of variance (ANOVA). First, the judgments showed increasing underconfidence across the trials,  $F(1, 59) = 66.5$ ,  $MSE = 183.1$ ,  $p < .05$ , effect size ( $ES$ ) = 1.1,

<sup>1</sup> In the 3 experiments reported here, response latency (the measure of retrieval fluency) was the time from the onset of the item cue (i.e., presentation of the first word of the pair for recall) until participants typed their responses and pressed the “Enter” key. Although responses were one-word nouns and rather short, within-individual differences in typing time across responses may have resulted in a somewhat less reliable measure of retrieval latency. Accordingly, we conducted a follow-up experiment (to be reported in detail elsewhere), in which participants responded by pressing the spacebar once they had retrieved a response. Responses could not be entered until the spacebar had first been pressed. Time to press the spacebar and time to enter the entire response were both recorded and were highly related (mean intra-individual  $r = .90$ ,  $SEM = .02$ ,  $p < .05$ ), which validates the measure used in the experiments reported here. As important, all of the outcomes reported here were replicated in the follow-up experiment using either measure of response latency.

Table 1  
*Absolute Accuracy of Judgments of Learning (JOLs) Across Trials*

Kind of JOL	Trial		Shift to underconfidence
	1	2	
Experiment 1			
Immediate	11.2 (3.1)	-12.3 (2.7)	-23.4 (2.5)
Delayed	-0.2 (2.0)	-5.2 (1.6)	-5.1 (1.6)
Experiment 2			
Immediate	14.1 (3.2)	-15.2 (2.9)	-29.3 (2.5)
Delayed	-7.2 (1.9)	-14.9 (2.4)	-7.7 (1.8)
Experiment 3			
Immediate	7.2 (4.0)	-20.6 (2.6)	-27.8 (4.3)
Delayed	-2.9 (1.9)	-8.7 (3.3)	-5.8 (2.8)

*Note.* Cell entries are difference scores between judgment of learning (JOL) magnitude and recall performance. Entries in parentheses are the corresponding standard errors of the mean. Both kinds of judgment in Experiment 2 and the delayed JOLs of Experiment 3 included pre-JOL recall attempts.

which replicates the UWP effect for both kinds of JOL. Second, the magnitude of the UWP effect was larger for immediate than delayed JOLs,  $F(1, 59) = 59.8$ ,  $MSE = 84.7$ ,  $p < .05$ ,  $ES = 1.7$ .

#### *Relations Between Retrieval Latency and Judgments*

We computed two gamma correlations (one for immediate JOLs and one for delayed JOLs) between each participant's correct-retrieval latencies on Trial 1 and the corresponding JOLs made on Trial 2. Means across participant's correlations were  $-.16$  ( $SEM = .07$ ) for immediate JOLs and  $-.03$  ( $SEM = .10$ ) for delayed JOLs,  $t(37) = 0.80$ ,  $p > .10$ . Although this correlation was not reliably different from 0 for delayed JOLs,  $t(44) = 0.33$ ,  $p > .10$ , it was for immediate JOLs,  $t(44) = 2.27$ ,  $p < .05$ . The latter outcome is consistent with the hypothesis that Trial 2 immediate JOLs are based partly on retrieval fluency during Test Trial 1.

#### *Linking Retrieval Fluency to the UWP Effect*

Evidence from Experiment 1 implicates retrieval fluency as a possible contributor to the UWP effect, because a negative correlation was evident between correct response latency and immediate JOLs. Even though this outcome suggests that retrieval fluency is a basis for immediate JOLs and hence could contribute to the UWP effect, it does not directly demonstrate that slower retrieval is related to increased underconfidence on Trial 2.

Accordingly, we further explored the retrieval-fluency hypothesis by directly linking differences in response latency on Trial 1 recall to underconfidence on Trial 2. This analysis was modeled from those used by Benjamin et al. (1998) and was conducted only for immediate JOLs, because only these judgments demonstrated a reliable negative correlation between retrieval fluency and JOLs. For each participant, items that had been correctly recalled prior to making JOLs were separated into three bins according to the latency of correct response (via Vincentizing per individual par-

ticipant). For immediate JOLs, latencies from correct recall on the Trial 1 test trials were used. For each bin, we computed the underconfidence on Trial 2 that contributed to the total underconfidence across all items on Trial 2.<sup>2</sup> This analysis provided an estimate of the amount of total underconfidence on Trial 2 that was attributed to items that had the fastest to slowest response latencies. According to the retrieval-fluency hypothesis, underconfidence on Trial 2 will be largest for those items in which responses had been retrieved most slowly on Trial 1.

The mean underconfidence for each latency bin is presented in Table 2, along with outcomes from a one-way ANOVA. Moreover, for interested readers, the mean JOL and mean recall performance for each bin (which in part were used to derive the critical measures in Table 2, see Footnote 2) are reported in Table B1 in Appendix B. What is striking from Table 2 is that underconfidence on Trial 2 did not increase as a function of increasing response latencies for correctly recalled items. In fact, underconfidence was evident even for responses that had been quickly retrieved and at a magnitude comparable with the underconfidence observed for items in the other latency bins. This outcome indicates that even though retrieval fluency was related to immediate JOLs in Experiment 1, it did not substantively contribute to the UWP effect.

#### Experiment 2

In Experiment 2, we further investigated the contribution of retrieval fluency to the UWP effect. Experiment 1 provided evidence that retrieval fluency may influence immediate JOLs, as suggested by the negative correlation between response latencies and immediate JOLs. By contrast, retrieval fluency did not appear to influence delayed JOLs, in that retrieval latencies for correctly recalled items during Trial 1 recall were not reliably correlated with delayed JOLs. We found this particular outcome quite surprising, because multiple hypotheses claim that delayed JOLs are based on retrieval fluency for correctly retrieved responses (e.g., Nelson & Dunlosky, 1991; Nelson, Narens, & Dunlosky, 2004). According to these hypotheses, however, and in contrast to our measure of retrieval fluency in Experiment 1, retrieval fluency relevant to delayed JOLs pertains more specifically to the outcome of a covert retrieval attempt that occurs immediately prior to making a delayed JOL. For instance, when shown a cue for a delayed JOL (e.g., "daffodil - ?"), an individual presumably attempts to retrieve the response (i.e., blood) and then bases the JOL on the success and fluency of that retrieval attempt (Koriat & Ma'ayan, 2005; Nelson et al., 2004; Nelson, Scheck, Dunlosky, & Narens, 1999). As illustrated in Figure 1B, this pre-JOL retrieval

<sup>2</sup> To estimate these values, for each participant we computed the difference score (between mean JOLs and mean recall) for each bin of items and multiplied this value by the proportion of items within each bin. Thus, the difference scores reported in Table 2 represent the contribution of each bin to the overall underconfidence from correctly recalled items on Trial 1. By adding the values within each row (across bins) in Table 2, one obtains the amount of the total UWP effect (see Table 1) that is attributable to correctly recalled items on Trial 1.

Table 2  
Underconfidence on Trial 2 for Correctly Recalled Items Grouped by Response Latency

Kind of JOL	Response latency bin			F	df	MSE
	Fastest	Middle	Slowest			
Experiment 1						
Immediate	-2.1 (0.50)	-1.7 (0.47)	-1.7 (0.42)	1.84	2, 96	1.43
Experiment 2						
Delayed	-5.2 (0.79)	-5.3 (0.78)	-5.5 (0.82)	0.20	2, 118	6.69
Experiment 3						
Immediate	-1.5 (0.43)	-1.9 (0.34)	-1.7 (0.40)	0.40	2, 62	1.69
Delayed	-3.4 (1.00)	-3.4 (1.10)	-2.8 (1.10)	1.22	2, 68	3.26

Note. For immediate judgments of learning (JOLs), response latency is the time to respond during Trial 1 recall (correct responses only). For delayed JOLs, response latency is the time to respond during Trial 2 pre-JOL recall (correct responses only). See text for detailed rationale behind these analyses. All *p* values for corresponding *F*s in Table 2 were greater than .10. Values in parentheses are standard errors of the means.

attempt is proximal to delayed JOLs made on Trial 2. Retrieval fluency of the prejudgment recall attempt may influence JOLs in a way that contributes to the UWP effect.

Thus, we evaluated the retrieval-fluency hypothesis again, but instead obtained a measure of retrieval fluency that occurred immediately prior to each delayed JOL. To do so, we modified the procedure of Experiment 1 as follows. For both kinds of JOL, participants were instructed to make an explicit pre-JOL recall

attempt immediately prior to making each JOL (for detailed discussion of this procedure, see Nelson et al., 2004). If retrieval fluency does influence delayed JOLs, response latencies during pre-JOL recall will negatively correlate with delayed JOLs. As important, assuming this relationship is found, if retrieval fluency also contributes to the UWP effect for delayed JOLs, then underconfidence on Trial 2 will be greater for items that had been the most slowly retrieved during pre-JOL recall.

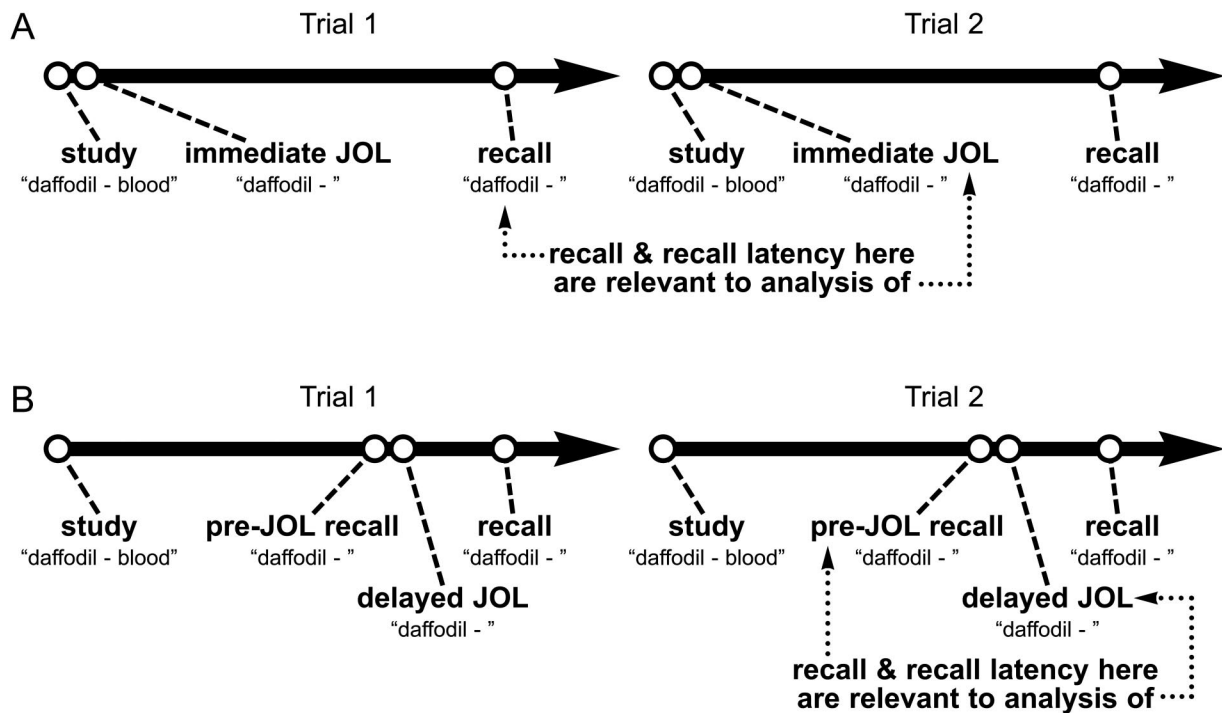


Figure 1. Appropriate analyses for evaluating the contribution of retrieval fluency to the UWP effect. Recall and recall latencies for Trial 1 recall attempts are most relevant for analyses of Trial 2 immediate judgments of learning (JOLs; A), whereas recall and recall latencies for Trial 2 pre-JOL recall are most relevant for analyses of Trial 2 delayed JOLs (B).



### Method

#### Participants, Design, and Material

The materials used in Experiment 2 were the same as those used in Experiment 1. Sixty undergraduates from UNCG participated to partially fulfill course requirements. None of the participants in Experiment 2 had participated in Experiment 1.

#### Procedure

The procedure of Experiment 2 was the same as in Experiment 1 with the exception that participants made a pre-JOL recall attempt immediately before making each JOL. This recall attempt was prompted in an identical manner to criterion recall in that participants were shown the first word of a pair and were asked to recall the second word. Immediately after this pre-JOL recall attempt, the participant made a JOL for the same item.

### Results and Discussion

#### The UWP Effect

The UWP effect was estimated as in Experiment 1, and the corresponding values are presented in Table 1 (and Table A1, Appendix A). Both main effects and the interaction were reliable. Both kinds of JOL also demonstrated a UWP effect, with increased underconfidence on Trial 2 compared with Trial 1,  $F(1, 59) = 107.8$ ,  $MSE = 190.1$ ,  $p < .05$ ,  $ES = 1.3$ . Finally, the reliable interaction,  $F(1, 59) = 73.3$ ,  $MSE = 95.8$ ,  $p < .05$ ,  $ES = 1.9$ , indicated that the overall shift to underconfidence was larger for immediate than delayed JOLs. These outcomes replicate those reported in Experiment 1.

#### Relations Between Response Latency and Judgments

As in Experiment 1, we computed the relationship between retrieval latencies and JOLs. These analyses were conducted for each of the two kinds of recall trial: recall on Test Trial 1 and pre-JOL recall on Trial 2. As illustrated in Figure 1, each kind of recall is most appropriate to the analysis of one or the other (but not both) of the JOLs. For instance, recall attempts during Trial 1 are most relevant to the analysis of Trial 2 immediate JOLs (see Figure 1A), whereas Trial 2 pre-JOL recall attempts are most relevant to the analysis of Trial 2 delayed JOLs (see Figure 1B). Pre-JOL recall is less relevant to immediate JOLs because participants can readily recall most responses from short-term memory when making immediate JOLs (Nelson et al., 2004). As explained earlier, pre-JOL recall is proximal to delayed JOLs and is highly predictive of final recall on that trial. Even though each analysis is not equally relevant to both kinds of JOL, we report all analyses for completeness.

**Trial 1 recall.** As in Experiment 1, a gamma correlation was calculated between the retrieval latencies of each participant's correct recall attempts on Trial 1 and the corresponding Trial 2 JOLs. Means across participant's correlations were  $-.18$  ( $SEM = .09$ ) for delayed JOLs and  $.02$  ( $SEM = .08$ ) for immediate JOLs,  $t(29) = 1.61$ ,  $p > .10$ . Whereas the correlation was reliably different from 0 for delayed JOLs,  $t(34) = 2.05$ ,  $p < .05$ , it was not reliable for immediate JOLs,  $t(43) = 0.24$ ,  $p > .10$ . An explanation for the latter inconsistency across experiments involving immediate JOLs will be examined in Experiment 3.

**Trial 2 pre-JOL recall.** A gamma correlation was calculated between the retrieval latencies of each participant's correct pre-JOL recall attempts on Trial 2 and the corresponding Trial 2 JOLs. The mean correlation was  $-.34$  ( $SEM = .05$ ) for delayed JOLs and was  $-0.08$  ( $SEM = 0.03$ ) for immediate JOLs,  $t(49) = 5.34$ ,  $p < .05$ ,  $ES = 1.5$ . The correlation for delayed JOLs was reliably different from 0,  $t(52) = 7.0$ ,  $p < .05$ , and unexpectedly, the relatively small correlation for immediate JOLs was also reliable,  $t(54) = 2.38$ ,  $p < .05$ .

#### Linking Retrieval Fluency to the UWP Effect

As in Experiment 1, we further explored the contribution of retrieval fluency to the UWP effect by linking differences in response latency to underconfidence on Trial 2. We computed underconfidence on Trial 2 as a function of latency bins (for the fastest, middle, and slowest response latencies) for only delayed JOLs in Experiment 2, which showed a substantial relationship with retrieval latency that could potentially contribute to underconfidence. As shown in Table 2, underconfidence was evident even for responses that had been quickly retrieved and at a magnitude comparable with the underconfidence observed for items in the other latency bins. Thus, retrieval fluency did not substantively contribute to the UWP effect for delayed JOLs.

### Experiment 3

Evidence from Experiments 1 and 2 indicated that the UWP effect occurs for immediate and delayed JOLs. However, evidence was inconsistent with a retrieval fluency account of the UWP effect, especially for immediate JOLs, which showed the largest underconfidence on Trial 2. In Experiment 1, the correlation between response latencies on Trial 1 recall and immediate JOLs on Trial 2 was low ( $-.16$ ) albeit reliable and was even lower ( $-.08$ ) in Experiment 2. One possible explanation for these weak correlations is that pre-JOL recall had a reactive effect on immediate JOLs. Retrieving responses just prior to making immediate JOLs on Trial 2 may subtly shift participants' attention away from retrieval fluency that occurred during recall on Trial 1. Even in Experiment 1, covert pre-JOL recall attempts elicited by making delayed JOLs may have had similar reactive effects on participants' immediate JOLs. If so, the low correlation between Trial 1 retrieval latencies and Trial 2 immediate JOLs in Experiment 1 may underestimate the actual influence of retrieval fluency on immediate JOLs.

Motivated by this possibility, we conducted a third experiment in which judgment delay was manipulated between participants. Besides obtaining a better estimate of the contribution of retrieval fluency, another advantage of this design is that the outcomes will be more relevant to explanations of the original UWP effect described by Koriat et al. (2002) in which participants made only immediate JOLs. Because the most appropriate analysis for estimating the contribution of retrieval fluency differs for the two kinds of JOL (see Figure 1), we used the following procedure in Experiment 3. For immediate JOLs, the two study-test trials did not include pre-JOL recall attempts, with retrieval fluency being estimated from correct recall on Test Trial 1. For delayed JOLs, pre-JOL recall attempts were included, and retrieval fluency was estimated from correct recall on Trial 2 pre-JOL recall.

## Method

### Participants, Design, and Materials

The participants were 70 undergraduates from UNCG who participated in partial fulfillment of a class requirement. None participated in Experiment 1 or 2. The materials were 30 pairs of concrete nouns taken from the list of pairs used in Experiments 1 and 2. All participants studied and judged the same 30 pairs, but participants in the immediate-JOL group made only immediate JOLs, and participants in the delayed-JOL group made only delayed JOLs. Apple iMac computers recorded all responses and response latencies.

### Procedure

The participants were randomly assigned to either the immediate-JOL group or the delayed-JOL group, with the restriction that an equal number of participants be assigned to both groups. As in the two previous experiments, participants completed two study–test trials of paired associates. The order of the 30 pairs of nouns was randomized at the outset of the experiment for each participant and presented one at a time for 6 s each. After participants in the immediate-JOL condition studied each pair, the computer prompted them to make an immediate JOL (as in Experiment 1). Participants in the delayed-JOL group studied all 30 items individually for 6 s each. They then made a pre-JOL recall attempt and delayed JOL for each of the same 30 items (as in Experiment 2), with the prompts for pre-JOL recall and delayed JOLs occurring in the same order as the corresponding items had been presented for study. After all 30 pairs were studied and judged, participants worked on a distracter task for 5 min. The computer then tested the participants via paired-associates recall, with the order of items being randomized anew. Trial 2 followed the same procedure as Trial 1, except the order of presenting items for study and test was randomized anew before study.

## Results and Discussion

### The UWP Effect

Absolute accuracy is reported in Table 1. A 2 (kind of JOL)  $\times$  2 (trial) ANOVA revealed a main effect across trials,  $F(1, 65) = 44.2$ ,  $MSE = 213.5$ ,  $p < .05$ ,  $ES = .80$ , which indicated increasing underconfidence across the trials. Also, the magnitude of the UWP effect was larger for immediate than delayed JOLs,  $F(1, 65) = 19.0$ ,  $MSE = 213.5$ ,  $p < .05$ ,  $ES = .90$ .

### Relations Between Retrieval Latency and Judgments

*Trial 1 recall.* A gamma correlation was calculated between the retrieval latencies of each participant's correct recall attempts on Trial 1 and the corresponding immediate JOLs on Trial 2 (as per Figure 1, Panel A). Retrieval latencies on Test Trial 1 reliably correlated with immediate JOLs,  $M = -.24$ ,  $SEM = .10$ ,  $t(25) = 2.32$ ,  $p < .05$ .

*Trial 2 pre-JOL recall.* A gamma correlation was calculated between the retrieval latencies of each participant's correct pre-JOL recall attempts on Trial 2 and the corresponding Trial 2 delayed JOLs (as per Figure 1, Panel B). Retrieval latencies for pre-JOL recall reliably correlated with delayed JOLs,  $M = -.36$ ,  $SEM = 0.07$ ,  $t(22) = 4.91$ ,  $p < .05$ .

### Linking Retrieval Fluency to the UWP Effect

Underconfidence as a function of latency bins (see Table 2) was computed for immediate JOLs on the basis of response latency of

recall on Trial 1 and for delayed JOLs on the basis of response latency for pre-JOL recall on Trial 2. Consistent with the previous experiments, underconfidence was evident even for responses that had been quickly retrieved and at a magnitude comparable with the underconfidence observed for items in the other latency bins.

## General Discussion

The present study was conducted to evaluate the retrieval-fluency hypothesis for the UWP effect. Evidence from previous research implicated retrieval fluency as a potential contributor because retrieval fluency was negatively related to both immediate JOLs (Benjamin et al., 1998; Matvey et al., 2001) and delayed JOLs (Nelson et al., 1999). Moreover, Benjamin et al., among others (e.g., Kelley & Lindsay, 1993), have demonstrated how fluency of retrieval can negatively bias people's metacognitive judgments away from criterion performance. Such biases bring us to our primary question. Namely, to what degree does the reliance of JOLs on retrieval fluency contribute to the shift to underconfidence that occurs across trials?

The answer to this question, both for immediate and delayed JOLs, is "minimally, if at all." Although some correlations between response latency and JOLs were reliably negative, the magnitude of these latency–JOL correlations was consistently low. In Experiment 2, the correlation between response latencies and immediate JOLs was close to zero ( $-.08$ ), yet the shift to underconfidence on Trial 2 for immediate JOLs was larger in magnitude than in the other two experiments (see Table 1). When we conducted a more fine-grained decomposition of underconfidence on Trial 2 based on response latency (see Table 2), underconfidence did not reliably change as a function of response latencies for correctly recalled responses. If retrieval fluency contributed to the UWP effect as predicted, underconfidence would have been larger for items with the longest response latencies.

Given that response latencies were often negatively correlated with JOLs, why did retrieval fluency fail to contribute to underconfidence? One possibility is that for paired-associate recall of episodic associations, slow responding may have been a valid predictor of criterion recall on Trial 2. We evaluated this possibility for the critical cases by correlating retrieval latencies for responses correctly recalled prior to making JOLs with whether those responses were recalled on the criterion test on Trial 2 (for converging evidence, see Table B1 of Appendix B). The correlations for items slated with delayed JOLs were  $-.41$  ( $SEM = .08$ ,  $n = 39$ ,  $p < .05$ , Experiment 2) and  $-.56$  ( $SEM = .08$ ,  $n = 16$ ,  $p < .05$ , Experiment 3). For immediate JOLs, the correlations were  $.01$  ( $SEM = .21$ ,  $n = 11$ ,  $p > .10$ , Experiment 1), and  $-.28$  ( $SEM = .21$ ,  $n = 7$ ,  $p > .10$ , Experiment 3). For immediate JOLs, quite a few participants' values were indeterminate because of the high conditional values, yet the correlations in general suggested that longer response latencies were related to subsequent forgetting. These trends were enough to offset any possible contribution of retrieval fluency to the UWP effect.

These results appear inconsistent with those reported by Benjamin et al. (1998). More specifically, whereas we found that the slowest latencies of prejudgment retrieval were often predictive of lower levels of criterion recall, Benjamin et al. reported the opposite (see introduction for details). This inconsistency, however, is more apparent than real, given a critical difference between the tasks used in these experiments. For Benjamin et al., the prejudg-

ment task involved retrieval of semantic memories (i.e., general knowledge questions) followed by criterion retrieval of episodic memories (i.e., free recall of the answers without the questions present). According to Benjamin et al., “when participants answer the question initially, they are being guided by the question on a search through semantic memory. . . . The longer they spend on such a search, the more salient or elaborated the entry they create in episodic memory for the event of having searched for that answer. The more elaborate the [episode], the more easily it can be accessed on a later free-recall task” (p. 56). By contrast, in the present experiments as well as others investigating the UWP effect (Finn & Metcalfe, 2004; Koriat et al., 2002; Meeter & Nelson, 2003; Scheck & Nelson, 2005; Simon, 2003; Tiede, Lee, & Leboe, 2004), recall prior to judgments and subsequent criterion recall both involved retrieval of episodic memories. In these cases, slow retrieval on an initial trial may indicate that the sought-after episode was not well elaborated in memory (compared with quickly retrieved episodes), and hence it would be indicative of slightly poorer recall on subsequent trials.<sup>3</sup>

Because the retrieval-fluency hypothesis for the UWP effect did not fare well in the present context, we briefly considered another explanation for the effect, which was recently offered by Scheck and Nelson (2005). They proposed that the UWP effect results from the anchoring and adjustment of JOLs. A participant makes a JOL by choosing an initial—and typically intermediate—anchor point on the JOL scale and bases subsequent JOLs around this value. A reluctance to use JOL values too far from this anchor point causes a participant’s JOLs to be underconfident when performance is high, which occurs across multiple study–test trials. Although the present experiments were not constructed to potentially disconfirm this anchoring-and-adjustment hypothesis, this hypothesis can readily explain our relevant data presented in Table A1. Certainly, given its current—albeit limited—empirical success, this hypothesis deserves further consideration.

### Summary

In three experiments, we replicated the UWP effect for immediate JOLs. The effect was also extended to delayed JOLs, although it was substantially smaller in magnitude. As in previous research, outcomes supported the hypothesis that JOLs are based on the fluency of retrieving responses, but retrieval fluency did not contribute significantly to the UWP effect, which rules out a leading factor that inappropriately biases metacognitive judgments in other contexts. Although other promising hypotheses have been offered for the UWP effect (e.g., see Koriat et al., 2002; Scheck & Nelson, 2005), this intriguing bias on JOLs must wait till future research provides a well-accepted explanation.

<sup>3</sup> Although we chose to characterize differences among these experiments in terms of the system approach (episodic vs. semantic) used by Benjamin et al. (1998) to explain their results, the differences across experiments can also be readily explained by a process-oriented approach. For instance, Benjamin et al.’s task first involved retrieving the answer to a specific general information question and free recalling the answers in the next phase, and such differences in cue-based retrieval processes may have resulted in their dissociation. By contrast, our task involved paired-associate recall during both trials, which is more likely to produce commonalities in interitem differences in retrieval processes across trials. (We

thank A. Koriat for these observations.) It is important to note that regardless of whether one prefers a systems approach or a processing approach, both provide principled explanations for the different outcomes in Benjamin et al. and the studies reported here.

### References

- Benjamin, A., & Bjork, R. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.
- Benjamin, A., 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.
- Finn, B., & Metcalfe, J. (2004, November). *Multitrial judgments of learning*. Poster session presented at the 45th annual meeting of the Psychonomic Society, Minneapolis, MN.
- Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24.
- 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.
- Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*, 222–232.
- Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica*, *113*, 123–132.
- Nelson, T. O., & Dunlosky, J. (1991). When people’s judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The “delayed-JOL effect.” *Psychological Science*, *2*, 267–270.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press.
- Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised method for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Method*, *9*, 53–69.
- Nelson, T. O., Scheck, P., Dunlosky, J., & Narens, L. (1999, November). *What is the basis for judgments of learning (JOLs) for recallable items?* Paper presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA.
- 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.
- Simon, D. A. (2003, November). *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.
- Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.
- Thiede, K. W. (1999). The importance of monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin & Review*, *6*, 662–667.
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1024–1037.
- Tiede, H. L., Lee, M., & Leboe, J. P. (2004, November). *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.

## Appendix A

## Analyses of JOL Magnitude and Recall Performance

For the benefit of interested readers, we have included descriptive values for JOLs and recall across all three experiments. In Table A1, we present mean JOL magnitudes and criterion recall across all items for both trials, which correspond to the derived measures of absolute accuracy reported in Table 1. For each experiment, we conducted a 2 (trial)  $\times$  2 (kind of JOL)  $\times$  2 (measure; JOL vs. Recall) ANOVA. For clarity and brevity, we present only those outcomes that pertain specifically to the UWP effect—most notably, the interaction between trial and measure (which establishes the UWP effect) and the three-way interaction (which indicates a larger UWP effect for immediate than delayed JOLs). Most important, all outcomes were consistent with those based on the derived measures of absolute accuracy described in the text.

In Experiment 1, the Trial  $\times$  Measure interaction was reliable,  $F(1, 59) = 66.5$ ,  $MSE = 91.5$ ,  $p < .05$ ,  $ES = 1.8$ , and the three-way interaction was also statistically reliable,  $F(1, 59) = 59.9$ ,  $MSE = 42.4$ ,  $p < .05$ ,  $ES = 2.2$ , indicating a larger UWP effect for immediate JOLs than for delayed JOLs. In Experiment 2, a reliable Trial  $\times$  Measure interaction,  $F(1, 59) = 107.8$ ,  $MSE = 95.1$ ,  $p < .05$ ,  $ES = 2.3$ , established the presence of a reliable UWP effect. As evident from inspection of Table A1, the UWP effect was again larger for immediate than delayed JOLs, which was substantiated by a reliable three-way interaction,  $F(1, 59) = 73.3$ ,  $MSE = 47.9$ ,  $p < .05$ ,  $ES = 2.5$ . In Experiment 3, the Trial  $\times$  Measure interaction,  $F(1, 65) = 44.2$ ,  $MSE = 106.8$ ,  $p < .05$ ,  $ES = 1.7$ , and the three-way interaction were again reliable,  $F(1, 65) = 19.0$ ,  $MSE = 106.8$ ,  $p < .05$ ,  $ES = 1.2$ , indicating that the UWP effect was present and larger for immediate than delayed JOLs.

Table A1  
*Mean JOL and Mean Criterion Recall for Experiments 1–3*

Kind of JOL	Trial 1	Trial 2
Experiment 1		
Immediate		
JOL	46 (2.5)	55 (3.0)
Recall	34 (2.6)	68 (3.0)
Delayed		
JOL	36 (2.5)	65 (3.0)
Recall	36 (2.7)	70 (3.1)
Experiment 2		
Immediate		
JOL	43 (2.9)	44 (3.5)
Recall	29 (2.3)	59 (3.3)
Delayed		
JOL	25 (2.4)	52 (3.7)
Recall	33 (2.6)	66 (3.2)
Experiment 3		
Immediate		
JOL	39 (2.9)	46 (3.8)
Recall	32 (3.6)	67 (4.1)
Delayed		
JOL	31 (3.8)	60 (4.7)
Recall	34 (4.5)	69 (4.0)

*Note.* Median judgments of learning (JOLs) were calculated for each participant and then averaged across participants to yield the reported means. “Recall” refers to the percentage of items correctly recalled across participants. Values in parentheses are standard errors of the means.

(Appendixes continue)



## Appendix B

## Analyses of JOL Magnitude and Recall Performance on Trial 2 As a Function of Response Latencies on Trial 1

In Table B1, we present mean values corresponding to correctly recalled items for each of the latency bins, which correspond to the analyses reported in Table 2. Note that subtracting the mean JOL and mean recall from Table B1 will not equal the amount of underconfidence presented in Table 2. The reason for this discrepancy is that the values reported in Table B1 are not adjusted for the proportion of total items that occur in each latency bin. By contrast, the values in Table 2 have been adjusted so that they accurately reflect the amount of total underconfidence on Trial 2 (Table 1) that arises from items within each latency bin. Most important, however, in all but one case (Experiment 1, discussed below) the effects in Table B1 correspond to the analyses presented in the text; that is, both support the same conclusions.

In Experiment 1, we conducted a 2 (measure)  $\times$  3 (latency bin; fastest, middle, slowest) ANOVA. First, a main effect for measure was obtained,  $F(1, 47) = 24.9$ ,  $MSE = 556.2$ ,  $p < .05$ ,  $ES = 0.69$ , which highlights the UWP effect for these items. As evident from inspection of Table B1, however, the main effect of latency bin,  $F(2, 94) = 1.8$ ,  $MSE = 76.1$ ,  $p > .10$ , and the interaction,  $F(2, 94) = 44.2$ ,  $MSE = 82.8$ ,  $p > .10$ , were not reliable. The lack of an effect of latency bin on JOL magnitude was surprising, because it is discrepant with the reliable correlation—albeit small in magnitude,  $-.16$ —between response latency and JOLs. This discrepancy arises from differential participant thresholds for making JOLs and can be misleading because averaging across participants (as reported in Table B1) can obscure an individual's use of retrieval fluency as a basis for JOLs. For this reason, we highlighted the more valid intra-individual measures in the text, which were first computed at the level of individual participants. Note, however, that this minor discrepancy does not undermine the main conclusions of this research.

In Experiment 2, the 2 (measure)  $\times$  3 (latency bin) ANOVA revealed a main effect of measure,  $F(1, 59) = 56.4$ ,  $MSE = 1117.9$ ,  $p < .05$ ,  $ES = 0.97$ , and a main effect of latency bin,  $F(2, 118) = 23.4$ ,  $MSE = 160.0$ ,  $p < .05$ ,  $ES = 0.88$ . The interaction was not reliable,  $F(2, 118) = 1.7$ ,  $MSE = 95.8$ ,  $p > .10$ . As evident from inspection of Table B1, these results are consistent with the conclusion that response latency is negatively related to JOLs, and most important, that underconfidence was robust across all latency bins.

In Experiment 3, the 2 (kind of JOL)  $\times$  2 (measure)  $\times$  3 (latency bin) ANOVA revealed a main effect of measure,  $F(1, 64) = 27.5$ ,  $MSE = 851.5$ ,  $p < .05$ ,  $ES = 0.63$ , and a main effect of latency bin,  $F(2, 128) = 7.8$ ,  $MSE = 123.6$ ,  $p < .05$ ,  $ES = 0.48$ . No other effects or interactions approached reliability. In summary, the analyses of raw scores for JOL magnitude and recall performance confirm the conclusions based on the

Table B1  
Mean JOLs and Recall on Trial 2 for Correctly Recalled Items Grouped by Response Latency

Kind of JOL	Response latency bin		
	Fastest	Middle	Slowest
Experiment 1			
Immediate			
JOL	82 (3.4)	80 (3.6)	85 (2.7)
Recall	96 (2.1)	93 (3.1)	97 (2.1)
Experiment 2			
Delayed			
JOL	73 (3.7)	68 (3.9)	60 (4.0)
Recall	97 (1.1)	95 (1.3)	89 (2.0)
Experiment 3			
Immediate			
JOL	79 (3.8)	77 (4.5)	76 (4.2)
Recall	94 (3.2)	99 (0.6)	93 (3.6)
Delayed			
JOL	85 (4.0)	80 (4.7)	80 (4.4)
Recall	100 (0.0)	96 (1.6)	92 (2.0)

*Note.* For immediate judgments of learning (JOLs), response latency is the time to respond during Trial 1 recall (correct responses only). For delayed JOLs, response latency is the time to respond during Trial 2 pre-JOL recall (correct responses only). Values in parentheses are standard errors of the means.

derived measures presented in the text. Namely, underconfidence on Trial 2 was robust across latency bins, indicating retrieval fluency contributed minimally—if at all—to the UWP effect.

Received February 3, 2005

Revision received June 10, 2005

Accepted June 22, 2005 ■