Mem Cogn (2011) 39:827–838
DOI 10.3758/s13421-010-0065-9

Predicting memory performance under conditions of proactive interference: Immediate and delayed judgments of learning

Christopher N. Wahlheim

Published online: 30 December 2010

Abstract Four experiments examined the monitoring accuracy of immediate and delayed judgments of learning (JOLs) under conditions of proactive interference (PI). PI was produced using paired-associate learning tasks that conformed to variations of classic A–B, A–D paradigms. Results revealed that the relative monitoring accuracy of interference items was better for delayed than for immediate JOLs. However, delayed JOLs were overconfident for interference items, but not for items devoid of interference. Intrusions retrieved prior to delayed JOLs produced inflated predictions of performance. These results show that delayed JOLs enhance monitoring accuracy in PI situations, except when intrusions are mistaken for target responses.

Keywords Proactive interference · Judgments of learning · Delayed JOLs · Metacognition

A central issue in metacognition research is the accuracy with which individuals can monitor their own learning. One way to examine this ability is by asking people to predict their future memory performance during or after study by making judgments of learning (JOLs). Investigations of the accuracy of JOLs in predicting memory performance have relied heavily on materials such as unrelated word pairs or prose passages (see Dunlosky & Metcalfe, 2009). Results have shown that people can monitor their learning of these materials with moderate to high levels of accuracy (e.g., Koriat, 1997; Nelson & Dunlosky, 1991). However, little research has examined the accuracy with which people can monitor their learning when materials are intended to produce high levels of interference (e.g., Diaz & Benjamin, 2011; Maki, 1999; Metcalfe, Schwartz, & Joaquim 1993). How accurately can people monitor their learning under conditions of proactive interference (PI)?

PI refers to the impairment in memory for recently learned information by previous learning of related materials (for reviews, see Anderson & Neely, 1996; Crowder, 1976). Accurate monitoring of one's learning under conditions of PI is important because PI is thought to play a critical role in forgetting. A common theme in the metacognition literature is that monitoring accuracy is intimately linked to the control of future behaviors (e.g., Koriat, Ma'ayan, & Nussinson, 2006; Nelson & Narens, 1990; Son & Metcalfe, 2000). One notion is that accurate monitoring can lead to effective study behaviors, resulting in enhanced memory performance. Thus, discovering ways to optimize monitoring accuracy in interference situations may also lead to the discovery of means by which to reduce the effects of PI.

The primary aim of the present study was to examine the accuracy with which people could predict their memory performance under conditions of PI using immediate and delayed JOLs. JOLs made immediately following study are often less accurate than JOLs that occur after a delay (e.g., Nelson & Dunlosky, 1991), presumably due to differences in the bases used for each type of judgment. Most would agree that delayed JOLs are superior to immediate JOLs because the success of retrieval attempts is the primary basis for delayed JOLs (Metcalfe & Finn, 2008; Nelson & Dunlosky, 1991, 1992; Spellman & Bjork, 1992; but see Kimball & Metcalfe, 2003). In contrast, bases for immediate JOLs, such as the degree of associative relatedness or the encoding operations performed by participants, are likely to be less valid (e.g., Koriat, 1997). Can delayed JOLs enhance monitoring accuracy in PI situations?

C. N. Wahlheim (✉)
Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130, USA
e-mail: cnwahlheim@gmail.com

Previous research has examined the accuracy of metacognitive judgments in interference situations. For example, Metcalfe et al. (1993) examined the sensitivity of feeling-of-knowing judgments (FOKs) in various PI situations. In their experiments, participants learned paired associates in several versions of classic A–B, A–D paradigms. The amount of interference was manipulated by holding cues constant and varying the relatedness of targets between two presentations (i.e., A–B, A–B; A–B, A–B'; A–B, A–D). The primary finding was that FOKs made for items that initially resulted in retrieval failure tended not to differ across conditions for which cues remained constant and were higher than in a control condition that included cues presented only once (A–B, C–D). In contrast, memory performance tended to be better for repeated cue-target pairs than for pairs in which cues remained constant and targets were changed. These results provided evidence that the familiarity of cues was the primary basis for FOKs and that the retrievability of targets had little influence on judgments. However, the insensitivity of FOKs to differences in performance across interference conditions is unlikely to generalize to JOLs because all items receive JOLs and only items that result in failed retrieval attempts receive FOKs.

Consistent with this notion, Maki (1999) found that JOLs made for all studied items were sensitive to the effects of retroactive interference (RI). RI is similar to PI, with the primary difference being that recent learning has deleterious effects on memory for previously learned information, instead of the reverse. In her experiments, participants completed study–test trials that consisted of numerical cues and word targets (e.g., 261–*farmer*) in two consecutive lists. As in Metcalfe et al. (1993), the second list contained various amounts of overlap between cues and targets in the two lists, which produced different levels of interference. After studying the second list, the cues from the first list appeared individually, and participants made JOLs regarding how likely they would be to remember the corresponding targets. Results revealed that JOLs were lower for cues paired with different targets in each list than for cues paired with the same targets across lists. These results showed that participants were sensitive to the response competition produced by pairing multiple targets with the same cue.

In related work, Diaz and Benjamin (2011) directly investigated the effects of PI and release from PI on immediate JOLs. In their experiments, PI was created using variants of the A–B, A–D paradigm. PI was produced by holding cues constant and re-pairing targets across lists (e.g., A–B, A–C, A–D), and release from PI was produced by presenting new cue–target pairs following these lists (e.g., E–F). A JOL followed each study item. Their experiments provided tests of whether JOLs were sensitive to the familiarity of cues (e.g., Metcalfe et al., 1993) or to target competition (e.g., Maki, 1999), or whether participants used an analytic basis that included theories about the effects of interference. Results revealed that JOLs decreased in a manner consistent with memory performance across PI trials, but JOLs continued to decrease as memory performance increased on the release trial. These results provided evidence that participants based their judgments on theories that considered the global effects of PI and that their theories did not account for the cue-specific effects of PI. Furthermore, results replicated when participants received two consecutive trials that included buildup and release from PI, suggesting that experience with the task did not inform participants' theories.

In contrast to these findings, others have shown that experience with PI can educate metacognitive judgments. For example, Jacoby, Wahlheim, Rhodes, Daniels, and Rogers (2010), and Wahlheim and Jacoby (2011) examined whether participants could be educated about the effects of PI when given two study–test trials in A–B, A–D paradigms that included feedback at the time of test. Results from measures of self-allocated study time revealed that participants did not differentially allocate study time to control and interference items on initial trials, even though performance was worse for interference items at the time of test. However, participants devoted more time to interference than to control items on the second trial, indicating increased awareness of memory differences for interference and control items. These results showed that participants lacked awareness of PI due to insufficient experience, and that additional experience with PI tuned participants to its effects. Moreover, confidence judgments made at the time of test better discriminated between correct and incorrect responses on interference items on the second test trial, suggesting that participants became sensitive to the cue-specific effects of PI. The primary difference between the JOLs made in Diaz and Benjamin's (2011) experiments and the confidence judgments made in the experiments by Jacoby et al. (2010) and by Wahlheim and Jacoby (2011) was that participants could base their confidence judgments on retrieval attempts, whereas other cues or theories were used as bases for immediate JOLs. In this vein, delayed JOLs made on the basis of retrieval attempts might be more sensitive to PI effects than immediate JOLs.

In the present experiments, versions of mixed-list paired-associate learning tasks that conformed to A–B, A–D paradigms were used to produce PI. Two lists of cue–target pairs appeared consecutively, and the nature of the items was manipulated by varying the relationship between cues and targets between lists. Interference items consisted of cue–target pairs in which cues remained constant and targets were changed between the first and second lists (A–B, A–D). Performance on interference items was then

compared with that on control items (Experiment 1) for which cue–target pairs did not overlap between the lists (A–B, C–D) or with that on facilitation items (Experiments 2–4) for which cue–target pairs remained constant between lists (A–B, A–B). JOL queries appeared as cues without their targets. The timing of JOLs was manipulated between participants, with one group being queried immediately following each item during study (immediate JOLs) and the other making judgments in a separate phase following study (delayed JOLs). The purpose of blocking delayed JOLs in this manner was to increase participants' reliance on retrieval attempts as the primary basis for judgments.

Monitoring accuracy of immediate and delayed JOLs was assessed by comparing the difference between JOL magnitudes and recall performance (i.e., calibration) and by computing correlations between JOLs and memory performance at the item level (i.e., resolution). In contrast to previous studies that used materials devoid of interference to examine delayed JOL accuracy, delayed JOLs were not expected to enhance calibration for interference items. In a PI situation, retrieval of items learned on an initial list (i.e., intrusions) can result in participants incorrectly holding those items to be target responses with high levels of confidence. Consequently, retrieval of intrusions prior to delayed JOLs should have the effect of misinforming JOL magnitudes by inflating judgments relative to other unsuccessful retrieval attempts (cf. Krinsky & Nelson, 1985). However, relative JOL accuracy should still benefit from delayed JOLs, because retrieval attempts should be valid bases for judgments made on targets or errors other than intrusions. That is, the extent to which retrieval attempts are invalid bases for judgments depends on the rate of intrusions, and intrusions should occur only for a subset of items. Retrieval attempts on the remaining items should still be more valid than other bases typically relied upon when JOLs are queried immediately following study (cf. Koriat, 1997). These predictions were tested in the following experiments.

Experiment 1

Experiment 1 compared the accuracy of immediate and delayed JOLs in a classic PI paradigm. Related word pairs were included in control (A–B, C–D) and interference (A–B, A–D) conditions.

Method

Participants

Forty Washington University undergraduates participated in exchange for course credit or $10 per hour. The immediate and delayed JOL groups each contained 20 randomly assigned participants. All the participants were individually tested.

Design and materials

A 2 (item type: control vs. interference) × 2 (JOL: immediate vs. delayed) mixed design was used. Item type was a within-participants variable, and JOL was a between-participants variable. The materials consisted of 98 three-word sets that included one cue word (e.g., *frog*) and two associated targets (e.g., *prince*, *lizard*). The forward associative strength between the cues and targets was weak to moderate (range: .01–.05) according to the norms of Nelson, McEvoy, and Schreiber (1998). Both targets in each set were the same length.

The control and interference conditions conformed to A–B, C–D and A–B, A–D arrangements, respectively. A group of 90 critical sets and a group of 8 buffer sets were created from the original 98 three-word sets. These groups were further divided such that the 90 critical sets produced three groups of 30 sets and the 8 buffer sets produced two groups of 4 sets. Cues and targets in each group were equated for average length, word frequency (Balota et al., 2007), and strength of association from cue to target. For the three groups of critical sets, one group served in the A–B portion of the control condition and remained constant across experimental formats, and the 2 remaining critical sets were rotated through the second lists of the control and interference conditions (i.e., C–D and A–D). The responses that served as alternate responses in the first list and targets in the second list were also counterbalanced across formats. The two groups of buffer sets served in the primacy and recency portions of the study list. Each of these portions was comprised of two control items and two interference items. Buffers also served as practice items at the time of test and remained constant across formats.

The first list (A–B) contained 68 pairs (e.g., *frog–prince*) consisting of two groups of 30 critical items and 4 buffer items assigned to each of the control and interference conditions. These pairs appeared 3 times each, for a total of 204 presentations. In the second list (C–D or A–D), 68 pairs were presented once each. Half were control items that were new to the experiment (C–D), and the other half were interference items that contained the same cue as that in the first list, paired with a new target (e.g., *frog–lizard*). At the time of test, eight cue words (buffer items) followed by a question mark (e.g., *frog*–?) were presented during a practice phase, and the remaining 60 critical cues were presented next, during a final cued recall test.

Procedure

All the stimuli appeared in white, lowercase letters on a black background in the center of a computer screen. In the

first list, word pairs appeared 3 times each in a fixed random order, with the restriction that pairs from the same condition did not occur more than 3 times consecutively. Each pair was presented for 1.5 s, followed by a 500-ms interstimulus-interval (ISI). The instructions were to read each word pair aloud quickly.

In the second list, the presentation order of pairs was restricted in the same manner as in the first list. However, these pairs were presented for 4 s each, followed by a 500-ms ISI, and participants were told to study each pair for an upcoming memory test and to read each pair aloud. Participants were also told that the test would include the left member of each studied pair and that the task would be to recall the right member from the second list. They were encouraged to consider the nature of the upcoming memory test while studying these pairs.

In the immediate JOL group, participants were told that, after studying each pair, they would be asked to rate their confidence regarding the likelihood of remembering the right member of the pair when given the left member on a test that would occur approximately 5 min later. In the delayed JOL group, participants received the same study and JOL instructions as in the immediate JOL group. However, the study and JOL phases occurred in separate blocks. The presentation order was the same in each block to equate the number of intervening presentations between study and JOL for each item. The interval between JOL and test was also similar for each JOL group. However, note that this arrangement produced a longer retention interval between the study and test phases for the delayed than for the immediate JOL group.

At the time of JOL, the left member of each pair appeared with a question mark (e.g., *frog*–?), and the message "Likelihood of remembering?" was presented below with a scale that ranged from 0% to 100%. A rating of 0% indicated that participants would not remember the right member and would only be guessing at the time of test. In contrast, a rating of 100% indicated absolute certainty that they would remember the right member at the time of test. Participants were told that intermediate values (between 0% and 100%) represented varying levels of certainty and that they should use the full range of the scale to make their ratings as precisely as possible. Participants made their ratings by moving a marker across a sliding scale, using a computer mouse. Once they had selected a value, they moved on to the next item by clicking a box labeled "Continue."

Finally, at the time of test, the left member of pairs from the second list were presented individually, and participants were told to recall the right member. They were encouraged to give a response for every item but could pass if they could not think of any word related to the cue. Participants completed a practice test before the final test. Cue–question mark pairs appeared in the same manner as for JOL queries. Items remained on the screen until participants gave a response. An experimenter recorded the responses made aloud.

Results and discussion

In all the experiments, the significance level for all tests was set at $\alpha = .05$.

Recall performance Table 1 shows that the probability of correct recall was better for control than for interference items (.66 vs. .60), $F(1, 38) = 12.50$, $\eta_p^2 = .25$. In addition, recall was better in the immediate than in the delayed JOL group (.73 vs. .53), $F(1, 38) = 22.03$, $\eta_p^2 = .37$. This difference was likely due to there being more intervening events between study and test in the delayed JOL group. Finally, a trend indicated that the recall advantage for control over interference items was larger for the immediate than for the delayed JOL group. However, there was not a significant item type × JOL interaction, $F(1, 38) = 2.64$, $p = .11$, $\eta_p^2 = .07$. These results show that PI effects on recall tended to be larger for the immediate than for the delayed JOL group.

JOLs In contrast to the pattern of recall performance, JOLs tended to be higher for interference than for control items (Table 1). This difference was larger for the delayed JOL group. There was no effect of immediate versus delayed JOLs, $F < 1$, a significant effect of item type, $F(1, 38) = 17.46$, $\eta_p^2 = .32$, and a significant item type × JOL interaction, $F(1, 38) = 9.46$, $\eta_p^2 = .20$. The results suggest that the retrieval of intrusions prior to judgments may have inflated delayed JOLs.

To explore this possibility, intrusion rates for the immediate and delayed JOL groups were compared, and JOLs were conditionalized on responses produced at test. The idea was that if pre-JOL retrieval attempts produced intrusions for some items, this retrieval practice would produce a higher rate of intrusions in the delayed than in the immediate JOL group. In addition, participants might be more confident in intrusions because those responses are more accessible than other errors. The results revealed more intrusions for delayed than for immediate JOLs (.24 vs. .14), $t(38) = 3.11$, $d = 0.98$, with delayed JOLs being higher for interference items that eventuated in intrusions than for other errors (.60 vs. .38), $t(19) = 3.42$, $d = 1.10$. These results support the idea that intrusions retrieved prior to delayed JOLs contributed to the inflation of judgments.

Calibration Because differences in recall performance between JOL groups complicated the interpretation of calibration analyses, calibration differences were examined between item types within groups. Table 1 shows that immediate JOLs were more underconfident for control than

Table 1 Correct recall proportion, judgment-of-learning (JOL) magnitudes, and calibration as a function of item type and JOL: Experiments 1–4

	Correct Recall		JOL Magnitude		Calibration	
JOL	Immediate	Delayed	Immediate	Delayed	Immediate	Delayed
Experiment 1						
Control	.77 (.03)	.54 (.03)	.59 (.03)	.59 (.03)	-.18 (.03)	.05 (.02)
Interference	.69 (.03)	.51 (.04)	.61 (.03)	.69 (.03)	-.08 (.03)	.18 (.03)
Experiment 2						
Facilitation	.87 (.03)	.81 (.03)	.73 (.02)	.76 (.03)	-.14 (.03)	-.05 (.02)
Interference	.61 (.04)	.51 (.04)	.53 (.03)	.71 (.03)	-.08 (.04)	.20 (.03)
Experiment 3						
Facilitation	.85 (.03)	.75 (.03)	.74 (.04)	.74 (.03)	-.11 (.03)	-.01 (.03)
Interference	.51 (.05)	.44 (.05)	.58 (.04)	.68 (.03)	.07 (.06)	.24 (.04)
Experiment 4						
Facilitation	–	.80 (.03)	–	.72 (.03)	–	-.08 (.02)
Interference	–	.51 (.04)	–	.63 (.03)	–	.12 (.03)

Note: Standard errors of the means are presented in parentheses

for interference items, $t(19) = 5.06$. In contrast, delayed JOLs were more overconfident for interference than for control items, $t(19) = 5.82$. All calibration scores were significantly different from zero, $t(19)$s > 2.08. As was expected, calibration of delayed JOLs was poorer for interference than for control items.

Resolution Resolution (Table 2) was assessed by computing mean within-participant gamma correlations between JOLs and recall performance (for a detailed rationale for using gammas, see Nelson, 1984). In contrast to calibration, scale differences do not affect relative JOL accuracy (when there are not ceiling or floor effects), thus allowing for the comparison of resolution across JOL groups. As was expected, delaying JOLs enhanced resolution. Resolution was better for delayed than for immediate JOLs (.80 vs. .12), $F(1, 38) = 182.61$, $\eta_p^2 = .83$, and for control than for interference items (.51 vs. .41), $F(1, 38) = 4.48$, $\eta_p^2 = .11$. There was no significant item type × JOL interaction, $F(1, 38) = 1.16$, $p = .29$, $\eta_p^2 = .03$. The finding that resolution for interference items was higher for delayed than for immediate JOLs indicates that retrieval attempts were a more valid basis on which to make judgments than were the bases used for immediate JOLs. However, the advantage of control over interference items suggests that retrieval attempts resulting in intrusions were invalid bases.

Experiment 2

The results from Experiment 1 showed that delayed JOLs could enhance monitoring accuracy in a PI situation. However, only the resolution measure revealed these benefits, whereas calibration estimates for interference items in the delayed JOL group were highly overconfident. Although this result was expected, the greater familiarity of cues in the interference than in the control condition may have contributed to this overconfidence. In Experiment 2, the familiarity of cues was matched for interference and noninterference item types by replacing control items with facilitation items in which cues and targets were repeated across lists. This change equated the contribution of cue familiarity across conditions and provided a more precise

Table 2 Within-participant gamma correlations between judgments of learning (JOLs) and recall performance as a function of item type and time of JOL: Experiments 1–4

	Time of JOL	
	Immediate	Delayed
Experiment 1		
Control	.15 (.05)	.88 (.05)
Interference	.10 (.05)	.73 (.05)
Experiment 2		
Facilitation	.42 (.08)	.87 (.08)
Interference	.16 (.05)	.67 (.06)
Experiment 3		
Facilitation	.43 (.10)	.67 (.10)
Interference	.15 (.06)	.62 (.06)
Experiment 4		
Facilitation	–	.88 (.03)
Interference	–	.62 (.04)

Note: Standard errors of the means are presented in parentheses

examination of the correspondence between JOLs and recall performance. Experiment 2 also examined whether the results from Experiment 1 would replicate and extend to another PI situation.

Method

Participants

Forty Washington University undergraduates participated in exchange for course credit or $10 per hour. The immediate and delayed JOL groups each contained 20 randomly assigned participants. All the participants were individually tested.

Design, materials, and procedure

The design, materials, and procedure were identical to those in Experiment 1, except that a 2 (item type: facilitation vs. interference) × 2 (JOL: immediate vs. delayed) mixed design was used. Facilitation items containing the same cue–target pairs in each of the two lists (A–B, A–B) replaced the control items.

Results

Recall performance Table 1 shows that recall performance was better for facilitation than for interference items (.84 vs. .56), $F(1, 38) = 171.61$, $\eta_p^2 = .82$. Recall also tended to be better in the immediate than in the delayed JOL group (.74 vs. .66), $F(1, 38) = 3.39$, $p = .07$, $\eta_p^2 = .08$. The item type × JOL interaction was not significant, $F < 1$.

JOLs Table 1 shows that JOLs were higher for facilitation than for interference items (.75 vs. .62), $F(1, 38) = 80.00$, $\eta_p^2 = .68$, and for delayed than for immediate JOLs (.74 vs. .63), $F(1, 38) = 7.47$, $\eta_p^2 = .16$. More important, a significant item type × JOL interaction, $F(1, 38) = 25.82$, $\eta_p^2 = .41$, revealed that the difference between facilitation and interference items was greater for immediate than for delayed JOLs. This interaction suggests that cue familiarity played little, if any, role in the overconfidence of delayed JOLs on interference items. Rather, retrieval of intrusions prior to judgments likely gave rise to this overconfidence.

Consistent with this possibility, follow-up analyses revealed a higher intrusion rate for the delayed than for the immediate JOL group (.30 vs. .18), $t(38) = 3.64$, $d = 1.15$. Also, delayed JOLs were greater for interference items eventuating in intrusions than for other errors (.64 vs. .31), $t(19) = 6.93$, $d = 1.96$. These results replicate Experiment 1 and further support the idea that intrusions held with high confidence were responsible for the inflation of delayed JOLs.

Calibration As in Experiment 1, calibration was compared for item types within JOL groups (Table 1). Results revealed that immediate JOLs tended to be more underconfident for facilitation than for interference items, $t(19) = 1.80$, $p = .09$, $d = 0.37$. In contrast, delayed JOLs were underconfident for facilitation items and overconfident for interference items. All calibration scores were significantly different from zero, $t(19)$s > 2.49, with the exception that the score for interference items in the immediate JOL group was marginally different from zero, $t(19) = 1.98$, $p = .06$. Again, the calibration of delayed JOLs was better in a condition devoid of interference (i.e., facilitation) than in an interference condition.

Resolution Resolution again benefited from delayed JOLs (Table 2). Five participants had perfect recall in at least one condition, which precluded them from resolution analyses. The results revealed that resolution for the remaining 35 participants was higher for delayed than for immediate JOLs (.77 vs. .29), $F(1, 34) = 55.72$, $\eta_p^2 = .62$, and higher for facilitation than for interference items (.64 vs. .42), $F(1, 34) = 11.07$, $\eta_p^2 = .25$. There was not a significant item type × JOL interaction, $F < 1$.

Summary In sum, the results from Experiment 2 replicated the results from Experiment 1. The magnitudes of immediate JOLs better reflected differences in recall performance between the facilitation and interference conditions than did those of delayed JOLs. In contrast, resolution was better for delayed than for immediate JOLs. These results showed that delayed JOLs continued to enhance resolution of interference items when facilitation, rather than control, pairs were included in the list. More important, the primary mechanism implicated in the inflation of delayed JOLs was the retrieval of intrusions.

Experiment 3

Experiments 1 and 2 demonstrated that delayed JOLs enhanced monitoring accuracy of interference items in terms of resolution but did not improve calibration scores as compared with noninterference items. The evidence presented thus far indicates that pre-JOL recall of intrusions likely drove the overconfidence in delayed JOLs for interference items. Experiment 3 tested this account by examining calibration as a function of the number of responses that participants judged as paired with cues prior to JOLs. It seems reasonable that the overconfidence in delayed JOLs for intrusions should be greater for interference items for which intrusions were the only response to come to mind, as compared with items for which more than one response came to mind. The idea is that retrieval practice of an intrusion alone will hurt memory

performance more than will retrieval practice of multiple responses that may include the target response. Furthermore, JOLs should be reasonably high in both situations, because the accessibility of intrusions is likely to be higher than for other errors (cf. Koriat, 1993, 1995). Consequently, the prediction was that delayed JOLs for interference items judged as paired with only one response would be more overconfident than JOLs for items judged as paired with two responses.

Method

Participants

Forty Washington University undergraduates participated in exchange for course credit or $10 per hour. The immediate and delayed JOL groups each contained 20 randomly assigned participants. All the participants were individually tested.

Design, materials, and procedure

The design, materials, and procedure were identical to those in Experiment 2, with the exception that participants made response judgments indicating whether one or two responses was paired with a cue prior to making each JOL. The instructions correctly informed participants that all cues appeared with either one or two responses across lists. For both JOL types, boxes labeled "1" and "2" appeared below the JOL prompts. Participants made their judgments by clicking the mouse cursor on either box. After choosing a response, participants clicked a box labeled "Continue" to move on to the JOL portion of the trial.

Results

Recall performance Table 1 shows that recall was again better for facilitation than for interference items (.80 vs. .48), $F(1, 38) = 222.76$, $\eta_p^2 = .85$. There was also a nonsignificant trend indicating that recall was better in the immediate than in the delayed JOL group (.68 vs. .60), $F(1, 38) = 2.37$, $p = .13$, $\eta_p^2 = .06$. There was no significant item type × JOL interaction, $F < 1$.

JOLs Table 1 also shows that JOLs were greater for facilitation than for interference items (.74 vs. .63), $F(1, 38) = 46.78$, $\eta_p^2 = .55$, and JOLs were not significantly different for immediate and delayed groups (.66 vs. .71), $F(1, 38) = 1.09$, $p = .30$, $\eta_p^2 = .03$. However, a significant item type × JOL interaction, $F(1, 38) = 9.21$, $\eta_p^2 = .20$, indicated that the difference between JOLs for facilitation and interference items was larger for the immediate than for the delayed JOL group.

A one-tailed *t*-test revealed that intrusions were again higher for delayed than for immediate JOLs (.29 vs. .22), $t(38) = 1.68$, $d = 0.53$, and delayed JOLs were greater for interference items eventuating in intrusions than for other errors (.67 vs. .38), $t(19) = 7.44$, $d = 1.41$. Together, these results again support the notion that pre-JOL retrieval of intrusions drove the overconfidence on interference items in the delayed JOL group.

Calibration The method of assessing calibration was the same as that in Experiments 1 and 2 (see Table 1). Results revealed that immediate JOLs for facilitation items were underconfident, $t(19) = -3.84$, whereas JOLs for interference items were well calibrated, $t(19) = 1.21$, $p = .24$. In contrast, delayed JOLs for facilitation items were well calibrated, $t < 1$, whereas JOLs for interference items were overconfident, $t(19) = 6.32$. The results again show well-calibrated delayed JOLs in a learning situation devoid of interference and inflated JOLs in a PI situation.

Resolution Replicating Experiments 1 and 2, resolution was again better for delayed than for immediate JOLs (see Table 2). Resolution analyses included 36 participants who did not have perfect recall in any condition. Resolution was higher for delayed than for immediate JOLs (.65 vs. .29), $F(1, 35) = 18.77$, $\eta_p^2 = .35$, and for facilitation than for interference items (.55 vs. .38), $F(1, 35) = 4.35$, $\eta_p^2 = .11$. There was not a significant item type × JOL interaction, $F(1, 35) = 1.84$, $p = .18$, $\eta_p^2 = .05$. These results again suggest that retrieval attempts prior to delayed JOLs were valid for most items.

Response judgments The accuracy with which participants could identify the number of responses paired with a cue was examined as a function of item type and JOL. Participants were extremely accurate in identifying the number of responses paired with facilitation items in both the immediate and delayed JOL groups. Facilitation items were more often judged as having been presented with one rather than two responses (.90 vs. 10), $F(1, 38) = 612.72$, $\eta_p^2 = .94$, and this effect did not differ for immediate and delayed JOLs, $F < 1$.

Of primary relevance to the claim that intrusions retrieved prior to delayed JOLs on interference items inflated judgments were the response judgments for interference items. The prediction was that participants' response judgments would be more accurate for interference items in the immediate than in the delayed JOL group because more forgetting would occur in the latter group. This is because forgetting of List 1 responses could produce errors on response judgments in the immediate JOL group, whereas forgetting of List 1 or List 2 responses could both produce errors in the delayed JOL group. Results revealed

that interference items were judged as having been paired with two responses more often in the immediate than in the delayed JOL group (.82 vs. .60), $t(38) = 4.22$, $d = 1.34$. These results show that the immediate JOL group recognized which interference items had been paired with two responses more accurately than did the delayed JOL group.

The role of such recognition failures in the inflation of JOLs was examined more closely by comparing JOLs with recall performance as a function of response judgments. The prediction was that the difference between predicted and actual recall performance would be greater for interference items judged as paired with one rather than two responses, because intrusions retrieved as the only response should drive final recall performance down more than when two responses come to mind, for reasons described above. In contrast, the timing of immediate JOLs should largely preclude retrieval of intrusions, resulting in little, if any, difference between JOLs and recall for either response judgment.

Consistent with these predictions, Fig. 1 shows that immediate JOLs for interference items did not differ from recall performance for either response judgment. In contrast, delayed JOLs were more overconfident for items judged as paired with one response, rather than with two responses. Separate 2 (measure: JOL vs. recall) X 2 (response judgment: 1 vs. 2) ANOVAs conducted for immediate and delayed JOL groups confirmed these findings. There were no significant effects in the immediate JOL group, Fs ≤ 1.39, ps $\geq .25$, η_p^2s $\leq .07$. However, for the delayed JOL group, results from 19 participants who made both response judgments for interference items revealed significant effects of measure, $F(1, 18) = 89.55$, $\eta_p^2 = .83$, and response judgment, $F(1, 18) = 26.25$, $\eta_p^2 = .59$, and a significant measure × response judgment interaction, $F(1, 18) = 19.94$, $\eta_p^2 = .53$.

A further possibility is that intrusions produced prior to delayed JOLs were likely to have been mistaken for target responses more often when only one response came to mind. To explore this possibility, intrusion rates for interference items were examined as a function of response judgment. Retrieval practice of intrusions was expected to increase the probability of their being recalled at test more when intrusions were retrieved alone than when other responses were also retrieved. Consistent with this prediction, the results showed that the intrusion rate for interference items was higher for items judged as having been paired with one response, rather than with two responses (.41 vs. .25), $t(18) = 3.26$, $d = 0.99$.

In addition, if intrusions were more likely to be mistaken for targets when only one response came to mind in the delayed JOL group, JOLs for these responses should not differ for responses that eventuated in intrusions and targets because the accessibility of both responses should be similar. In contrast, JOLs should be higher for eventual targets than for intrusions when two responses came to mind because there would be more information available on which to base judgments (e.g., effects of response competition, list membership of each candidate, etc.). Analyses included data from the 14 participants who produced targets, intrusions, and other errors at the time of test and made both response judgments on interference items in the delayed JOL group. For interference items judged as paired with one response, JOLs did not differ for eventual targets and intrusions (.74 vs. .76), $t(13) = 0.26$, $p = .80$, $d = 0.10$. In contrast, JOLs for items judged as paired with two responses were greater for eventual targets than for intrusions (.88 vs. .61), $t(13) = 3.47$, $d = 1.29$. These results support the idea that intrusions were more often mistaken for target responses for interference items judged as paired with one response, instead of two responses, in the delayed JOL group.

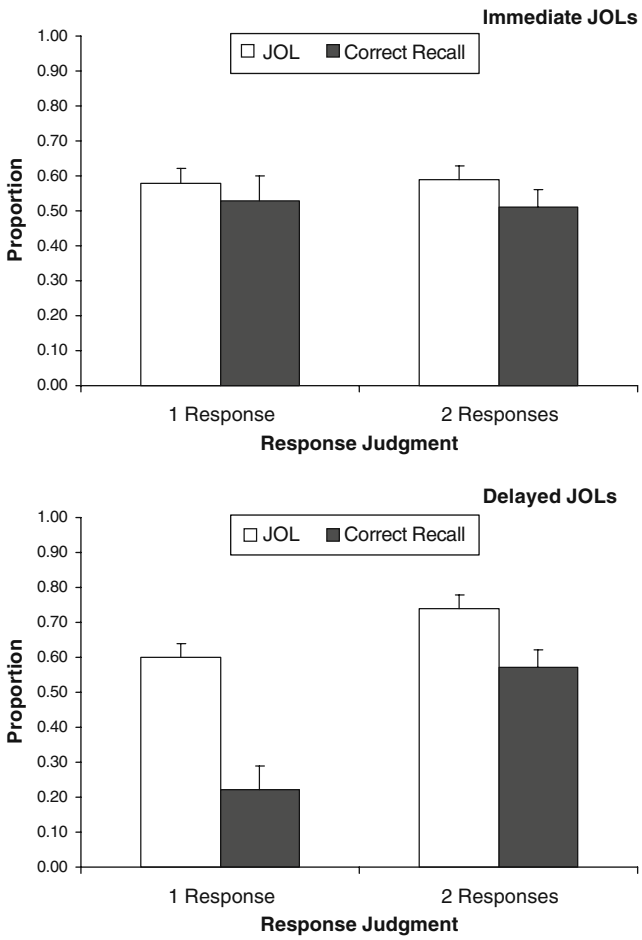

Fig. 1 Predicted and actual recall proportions for interference items as a function of response judgment in immediate and delayed judgment-of-learning (JOL) groups: Experiment 3

Summary The results of Experiment 3 provide further evidence suggesting that retrieving intrusions prior to

delayed JOLs inflates judgments. Retrieval practice with intrusions prior to judgments in the delayed JOL condition contributed to overestimations of recall performance in two ways. First, the accessibility of intrusions and target responses were sufficiently similar to lead participants to sometimes mistake intrusions for target responses. Second, retrieval practice of intrusions prior to delayed JOLs resulted in impaired recall performance due to an increase in intrusions at the time of test. Thus, inflated JOLs and impaired recall performance resulting from recall of intrusions prior to delayed JOLs gave rise to the overconfidence for interference items.

Experiment 4

The results from Experiments 1–3 strongly support the idea that retrieval of intrusions prior to delayed JOLs inflated judgments. In Experiment 4, direct evidence was sought for this notion by directly measuring retrieval attempts prior to delayed JOLs. Experiment 4 was the same as the delayed JOL group in Experiment 2, with the exception that participants made overt retrieval attempts prior to judgments. In addition, retrieval latencies for pre-JOL recall were recorded. The addition of overt retrieval attempts made it possible to verify that the retrieval of intrusions was responsible for the overconfidence of delayed JOLs for interference items, and allowed for the examination of resolution of interference items as a function of the type of response produced prior to JOLs. The expectation was that the pattern of overconfidence found in earlier experiments would replicate and that direct evidence for the role of pre-JOL intrusions in this overconfidence would obtain. Furthermore, resolution on interference items was not expected to differ from resolution on facilitation items when analysis did not include pre-JOL intrusions because the removal of intrusions would leave only items for which retrieval attempts produced targets or nonintrusion errors (i.e., valid bases for judgments).

Method

Participants

Twenty Washington University undergraduates participated in exchange for course credit or $10 per hour. All the participants were individually tested.

Design, materials, and procedure

The design, materials, and procedure were the same as those in Experiment 2, with the exception that participants made overt retrieval attempts prior to JOLs and the computer recorded their retrieval latencies. A box labeled "Click here to respond" appeared with the cues presented in the delayed JOL phase. Participants were told to click the mouse cursor on the box when the List 2 response came to mind, or when they were ready to guess if they could not remember the List 2 response. Next, participants made their responses aloud, and an experimenter recorded them.

Results

Recall performance As in Experiments 2 and 3, recall performance on the final test was better for facilitation than for interference items (Table 1), $t(19) = 12.46$, $d = 1.71$. Note that recall performance was nearly identical in Experiments 2 and 4. In addition, the rate of intrusions in Experiment 4 (.25) was comparable to those in the previous experiments. These results indicate that participants likely made covert retrieval attempts prior to delayed JOLs in the previous experiments even when not instructed to do so.

Pre-JOL recall Recall performance for retrieval attempts made prior to JOLs (i.e., pre-JOL recall) was nearly identical to performance on the final test. Pre-JOL recall was higher for facilitation than for interference items (.81 vs. .51), $t(19) = 12.42$, $d = 1.83$, and did not differ from final recall, $F < 1$. In addition, the proportion of responses that were the same for pre-JOL and final recall was extremely high and better for facilitation than for interference items (.96 vs. .91), $t(19) = 2.84$, $d = 0.81$. These results provide evidence that the responses produced on pre-JOL retrieval attempts and final recall were the same for most items. In addition, these results support the assumption in the previous experiments that intrusions made at the time of test often reflect intrusions made prior to JOLs. The proportion of items for which intrusions occurred on both pre-JOL and final recall was quite high (.88). Finally, the rate of intrusions did not differ for pre-JOL and final recall (.24 vs. .25), $t = 1.04$, $d = 0.14$.

Retrieval latencies for pre-JOL recall revealed that target responses were retrieved more quickly than intrusions, and that intrusions were retrieved more quickly than other errors (2,862 vs. 3,896 vs. 6,445 ms), $t(19)$s ≥ 2.38, ds ≥ 0.55. In addition, the correlation between retrieval time and JOLs computed across participants for interference items that produced intrusions during pre-JOL recall tended to be negative, $r = -.41$, $p = .08$. These results suggest that overconfidence in interference items increased with the accessibility of intrusions.

JOLs Table 1 shows that JOLs were higher for facilitation than for interference items, $t(19) = 5.19$, $d = 0.66$. In addition, JOLs were higher for intrusions than for other

errors for both pre-JOL recall (.62 vs. .26) and final recall (.60 vs. .27), $F(1, 19) = 48.76$, $\eta_p^2 = .72$. The latter results further support the notion that retrieval of intrusions inflated delayed JOLs.

Calibration Consistent with Experiment 2, Table 1 shows that delayed JOLs were underconfident for facilitation items, $t(19) = -3.88$, and overconfident for interference items, $t(19) = 4.05$. However, when interference items that produced pre-JOL intrusions were removed from the analysis, the results revealed that delayed JOLs (.62) did not differ from recall performance (.65), $t(19) = 1.12$. These results provide evidence that the retrieval of intrusions gave rise to the inflation of delayed JOLs.

Resolution One participant was excluded from resolution analyses due to perfect performance in the facilitation condition. Table 2 shows that resolution was again better for facilitation than for interference items, $t(18) = 5.22$, $d = 0.29$, resulting from overconfidence due to the retrieval of intrusions on interference items. An examination of Fig. 2 reveals further support for this conclusion, in that resolution did not differ between facilitation items and interference items that did not produce intrusions during pre-JOL recall, $t < 1$. In addition, resolution computed for interference items on which intrusions were retrieved prior to JOLs (for the 5 participants with variance in their performance) revealed a mean gamma correlation that was not significantly different from zero (-.21), $t < 1$. Taken together, the results from these experiments provide evidence that delayed JOLs can produce high levels of relative monitoring accuracy in a PI situation. This occurs when pre-JOL retrieval does not produce intrusions that are mistaken for target responses.

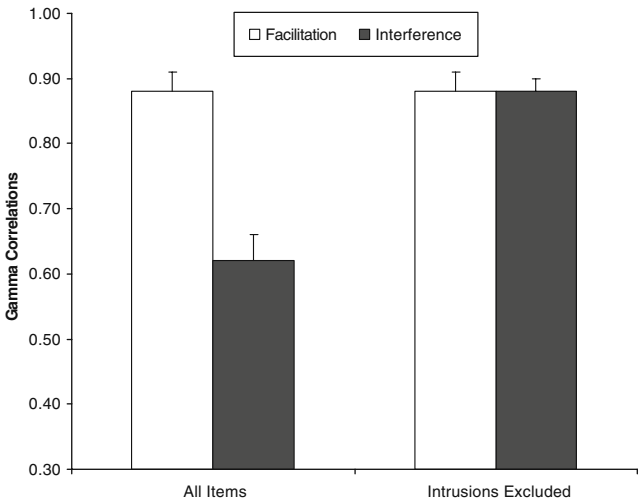

Fig. 2 Monitoring resolution for delayed judgments of learning (JOLs) as a function of item type and whether items that elicited intrusions prior to JOLs for interference items were included in the analysis: Experiment 4

General discussion

The primary aim of the present experiments was to determine the extent to which immediate and delayed JOLs predict memory performance under conditions of PI. The results showed that relative monitoring accuracy was better for delayed than for immediate JOLs; however, delayed JOLs were still overconfident for interference items. Further analyses revealed that the retrieval of intrusions prior to delayed JOLs inflated judgments, relative to those made on other errors. These results suggest that retrieval attempts can serve as both valid and invalid bases on which to make judgments in a PI situation and that the validity of this basis depends on the quality of retrieval.

Proactive interference and theories of delayed JOL effects

In their original study of delayed JOLs, Nelson and Dunlosky (1991) proposed that retrieval attempts from long-term memory made prior to delayed JOLs provide bases for JOLs that are more diagnostic of future recall performance than are immediate JOLs because the latter judgments are impaired by noise from short-term memory. According to the original conception of the monitoring dual memories (MDM) account, delayed JOLs should enhance both resolution and calibration for interference and noninterference items beyond that of immediate JOLs. This prediction holds for both types of items because noise from short-term memory should not influence retrieval attempts made prior to delayed JOLs for either item type. The results from the present experiments are incompatible with this account in that delayed JOLs produced overconfidence on interference items and did not improve resolution on interference items to the same extent as for control or facilitation items.

However, an auxiliary assumption of the MDM account not explicitly described in the original formulation of the theory is that people have difficulty evaluating the quality of memories that come to mind (J. Dunlosky, personal communication, November 30, 2010). Thus, retrieval from long-term memory is necessary for improving JOL accuracy, but the improvement depends on the quality of the retrieved information. If this assumption is considered, the results from the present experiments are compatible with the MDM account. Specifically, the results from Experiment 4 showed that when analyses of interference items did not include intrusions, delayed JOLs corresponded well with recall performance, and resolution did not differ for facilitation and interference items. These results are consistent with the

notion that delayed JOLs are more accurate than immediate JOLs because people can rely on the contents of long-term memory as a basis for judgments and that intrusions retrieved from long-term memory can hurt the accuracy of delayed JOLs.

Another plausible account of the delayed JOL advantage, the self-fulfilling prophecy hypothesis (Spellman & Bjork, 1992), posits that this effect occurs because of the retrieval practice prior to delayed JOLs. This retrieval practice artifactually boosts memory for the retrieved response, and, consequently, items that elicit responses receive higher JOLs than do items that do not elicit responses. The results in the present study are compatible with this account in two ways. First, intrusion rates were higher for delayed than for immediate JOLs, which was presumably due to retrieval practice of intrusions prior to judgments. Second, delayed JOLs that eventuated in intrusions received higher JOLs than did those that eventuated in other errors. This suggests that responses made more accessible through retrieval practice received higher JOLs.

In a similar vein, Koriat's (1993, 1995) accessibility hypothesis holds that there is a positive relationship between the accessibility of information stored in memory and the magnitude of metacognitive judgments. Indeed, Metcalfe and Finn (2008) have shown that the accessibility of responses is a primary basis for delayed JOLs. According to this account, the accuracy of delayed JOLs should be impaired in an interference situation since the ease with which a response comes to mind should be similar for target responses and intrusions mistaken for target responses retrieved prior to delayed JOLs. The results in Experiment 3 are in line with this notion, because the magnitude of JOLs for interference items judged as paired with one response did not differ between items that eventuated in targets or intrusions at the time of test. This was presumably due to participants misattributing the high accessibility of intrusions retrieved alone as an indication that those responses were from the target list.

Unintended interference effects and JOL accuracy

Interference might also play a role in JOL accuracy in experiments using materials that were not intended to produce high levels of interference. That is, it is unlikely that any experiments are completely free of interference. Even when not designed to produce interference, intralist and extralist intrusions are likely to occur sometimes because of unintended similarities among items, global interference, and so forth, and such interference might vary with factors such as list length. Undetected interference effects of this sort might commonly occur in JOL studies and might be partially responsible for the finding of only moderate correlations between JOLs and recall performance (cf. Koriat, 1997). Consideration of unintended interference effects may prove critical in discovering ways to optimize monitoring accuracy.

Diminishing the effects of proactive interference: Educating metamemory

As was mentioned in the introduction, finding ways to enhance awareness of PI is important for discovering ways to diminish its effects. Enhancing awareness of PI is of practical importance because methods that are successful in doing so may prove useful in training efforts aimed at the remediation of age-related memory deficits. In this vein, Jacoby et al. (2010) demonstrated that young and older adults could learn to diminish the effects of PI when given two trials with feedback at the time of test. Participants' heightened awareness of PI was thought to produce these effects, as evidenced by their more effective use of study time and reduced rates of intrusions held in high confidence on a second trial. Along with the results from the present study, these findings suggest that the influence of PI on memory performance will be strongest when there is low awareness of its effects. Consequently, training efforts might stand to benefit from procedures that highlight the subjective experience that accompanies items that produce intrusions that are mistaken for target responses.

Conclusion

Much has been made of the importance of optimizing monitoring accuracy to improve the control over future learning behaviors, and delayed JOLs are one means by which to do so. However, the finding that the benefits of delayed JOLs are limited in PI situations suggests that one should consider other means of optimization. Finding ways to optimize monitoring in interference situations is important because these situations are common in many real-world contexts (e.g., remembering where one last parked). Thus, the discovery of optimal bases for monitoring in these situations will be an important goal for future research.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (2nd ed., pp. 237–313). San Diego: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman,

R. (2007). The english lexicon project. *Behavior Research Methods, 39*, 445–459.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale: Erlbaum.

Diaz, M., & Benjamin, A. S. (2011). The effects of proactive interference (PI) and release from PI on judgments of learning. *Memory & Cognition.* doi:10.3758/s13421-010-0010-y.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks: Sage.

Jacoby, L. L., Wahlheim, C. N., Rhodes, M. G., Daniels, K. A., & Rogers, C. S. (2010). Learning to diminish the effects of proactive interference: Reducing false memory for young and older adults. *Memory & Cognition, 38*, 819–828.

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

Koriat, A. (1993). How do we know what we know? The accessibility model of feeling of knowing. *Psychological Review, 100*, 609–639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311–333.

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

Krinsky, R., & Nelson, T. O. (1985). The feeling of knowing for different types of retrieval failure. *Acta Psychologica, 58*, 141–158.

Maki, R. H. (1999). The roles of competition, target accessibility, and cue familiarity in metamemory for words pairs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1011–1023.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1084–1097.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851–861.

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

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

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

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgments-of-learning effect? *Psychological Science, 3*, 317–318.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation,* (*vol. 26,* pp. 125–173). New York: Academic Press.

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.

Spellman, B. A., & Bjork, R. A. (1992). People's judgments of learning are extremely accurate at predicting subsequent recall when retrieval practice mediates both tasks. *Psychological Science, 3*, 315–316.

Wahlheim, C. N., & Jacoby, L. L. (2011). Experience with proactive interference diminishes its effects: Mechanisms of change. *Memory & Cognition.* doi:10.3758/s13421-010-0017-4.

Author Note

This research was supported by Binational Science Foundation Grant 2005356 to Morris Goldsmith and Larry Jacoby. I express appreciation to John Dunlosky, Bridgid Finn, Larry Jacoby, Katherine Rawson, and Henry Roediger for their helpful comments and suggestions concerning this research. I thank Sarah Arnspiger, Emily Cokorinos, Lauren Guenther, and Dan Howard for their assistance with data collection, and Rachel Teune for her assistance with data collection and editing of this manuscript. Correspondence concerning this article should be addressed to Christopher N. Wahlheim, Department of Psychology, Washington University, St. Louis, MO 63130. E-mail: cnwahlhe@wustl.edu.