Supplemental Material A (Jang, Wallsten, and Huber): Complete Equations of the Stochastic Detection and Retrieval Model (SDRM)

Below, the SDRM is formalized with the assumption of linked criteria and with the assumption of independent criteria.

Let *h* be a bivariate normal distribution with *x* and *y* having a mean of zero and a standard deviation of one:

$$h(x, y, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right).$$

Let *i* be one of *n*+1 confidence response categories. The independent criteria model is derived by considering memory strength values (*y*) that are bracketed between two adjacent criteria, regardless of the order of the other criteria. For example, on a given JOL trial the criteria might be arrayed left to right C_1, C_3, C_2, C_4, C_5. In this case, a memory trace strength *y* that falls between C_3 and C_2 also is between C_1 and C_2 and between C_3 and C_4, so can lead to judgment J_1 or J_3. As a result, for the extreme confidence ratings ($i = 0$ or $i = n$), there is no difference between the models aside from the normalization factor that is necessary for the independent criteria model; in the equations below the extreme confidence values are shown with an equivalence relation although this should be a proportional relation for the independent criteria model. The proportional values of the independent criteria model are converted into a joint probability distribution by summing the *n*+1 proportional values across both recalled and not recalled responses and then dividing these proportional values by this total.

When $i = 0$,

$$p(J_i, \text{recalled}) = \iint h(x, y, \rho) N(x|C_M, \sigma_M)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) = \iint h(x, y, \rho)[1 - N(x|C_M, \sigma_M)][1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

when $0 < i < n$,

Linked version of the SDRM:

$$p(J_i, \text{recalled}) = \iint h(x, y, \rho) N(x|C_M, \sigma_M)[N(y|C_i, \sigma_C) - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) = \iint h(x, y, \rho)[1 - N(x|C_M, \sigma_M)][N(y|C_i, \sigma_C) - N(y|C_{i+1}, \sigma_C)]dxdy,$$

Independent version of the SDRM:

$$p(J_i, \text{recalled}) \propto \iint h(x, y, \rho) N(x|C_M, \sigma_M) N(y|C_i, \sigma_C)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) \propto \iint h(x, y, \rho)[1 - N(x|C_M, \sigma_M)] N(y|C_i, \sigma_C)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

and when $i = n$,

$$p(J_i, \text{recalled}) = \iint h(x, y, \rho) N(x|C_M, \sigma_M) N(y|C_i, \sigma_C)dxdy,$$

$$p(J_i, \text{not recalled}) = \iint h(x, y, \rho)[1 - N(x|C_M, \sigma_M)] N(y|C_i, \sigma_C)dxdy.$$

Supplemental Material B (Jang, Wallsten, and Huber): Complete Method and Results of the Experiment

Method

Participants. Two hundred and twenty-five undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation. Forty-five participants were assigned randomly to each of five groups.

Materials. Stimuli consisted of 60 concrete unrelated noun-noun pairs (Concreteness ≥ 6.10; norms from Paivio, Yuille, & Madigan, 1968). The first six study pairs of each list were excluded from data analyses, even if they were tested, so as to prevent primacy effects; and the last six were not even tested, so as to prevent recency effects: the remaining 48 pairs were the only ones analyzed.

Design. The experimental design was a 5 × 2 mixed factorial with type of practice (control, S, SJ, ST, and SJT) manipulated between subjects and JOL timing (immediate and delayed) manipulated within subjects. Word pairs were assigned randomly to the immediate- and delayed-JOL conditions for each participant and were randomly sequenced anew for each study, JOL, and test phase. Pairs receiving more than one JOL were consistently assigned to the immediate- or delayed-JOL conditions for all JOLs. Furthermore, one word of the pair was randomly assigned the role of cue and the other the role of target, and this remained the case even if the pair underwent more than one JOL or more than one cued-recall test. To describe the practice manipulation, we denote the cycle of study-JOL rating-recall as SJT. Control participants had no prior practice and went through just one SJT cycle. Participants in Condition S went through one study cycle, S, and then a complete SJT cycle. Those in Condition SJ went through one study-JOL cycle, SJ, and then a full SJT cycle. Those in Condition ST went through

one study-test cycle, ST, prior to SJT. And, finally, those in Condition SJT went through a full study-JOL-test cycle, SJT, and then did so a second time.

Procedure. All participants were instructed to study word pairs and to indicate their JOL for a pair whenever just one of the two words appeared (i.e., when just the cue word appeared). During the study phase, each pair appeared in the center of the screen for 5 s. Pairs destined for immediate and delayed JOLs underwent the same study procedure and appeared in the same study list in a randomly intermixed fashion. However, JOLs were elicited right after the offset of a study pair assigned to the immediate-JOL condition, whereas JOLs were elicited after all the pairs had been studied for a study pair assigned to the delayed-JOL condition. In both cases, participants were prompted with the cue word and asked, "How confident are you that in about ten minutes from now you will be able to recall the second word of the item when prompted with the first word?" The participants reported their estimate on a scale of "0 = *definitely will not recall*, 20 = *20% sure*, 40 = *40% sure*, 60 = *60% sure*, 80 = *80% sure*, and 100 = *definitely will recall*". JOL responses were self-paced.

The test phase was initiated immediately after the end of the study phase. The first 6 test pairs were the first 6 study pairs, although these were not analyzed. The next 24 test pairs corresponded to the 24 study pairs that were studied in positions 7 through 30 of the study list except that these appeared in a different randomly determined order during the test list. The next set of 24 test pairs corresponded to the 24 study pairs that were studied in positions 31 through 54 of the study list in a randomly determined order. The last 6 pairs of the study list were not tested. This procedure was used to make sure that there was a sufficiently long delay between initial study and final test. For participants in the S, SJ, ST, and SJT conditions, the same procedures were used for the second cycle of the study, JOL and test phases. During the second

cycle, the order of the study and test lists was a new random ordering of the words with the stipulation that each pair remained part of the same component of the study and test list (i.e., first 6, next 24, second set of 24, and last 6). During the test phase, in which recall was also self-paced, the participants typed the target word when cued by the first word of the pair. If they had no guess, they typed *NEXT* to proceed to the next test trial.

Results

Below, all statistical tests use $\alpha = .05$, and for those that are significant, we report effect size as partial eta squared (η_p^2).

JOL accuracy in all conditions. Figure S1 shows mean gamma as a function of JOL timing and type of practice. A two-way analysis of variance (ANOVA) showed that both main effects were significant, $F(4, 202) = 3.55$, $MSE = .06$, $\eta_p^2 = .07$, for type of practice; and $F(1, 202) = 226.92$, $MSE = .05$, $\eta_p^2 = .53$, for JOL timing. The interaction was also significant, $F(4, 202) = 7.02$, $MSE = .05$, $\eta_p^2 = .12$.

All gammas for delayed JOLs were close to ceiling. As a consequence, the five practice conditions under delayed JOL did not differ, $F(4, 202) < 1$. From another perspective, the delayed-JOL effect held regardless of the type of practice. Turning to the immediate-JOL conditions, applying Tukey's HSD test, the ordering of JOL accuracy was C ≈ S ≈ SJ < ST ≈ SJT. Applying Kolmogorov–Smirnov tests to the response probability distributions yielded the same pattern of results. A new finding is that the testing-JOL effect is due only to test practice and not to additional study or JOL rating practice.

JOL accuracy in Condition SJT. Condition SJT consists of two SJT cycles, which provides an opportunity for careful analysis of whether the second round of JOL ratings depend

on any aspects of the first cycle. We calculated gamma correlations across the two SJT cycles for each of the immediate- and delayed-JOL conditions (i.e., eight gamma correlations in total) as follows: (1) gamma between the first J and the first T (J_1T_1); (2) gamma between the second J and the second T (J_2T_2); (3) gamma between the second J and the first T (J_2T_1); and (4) gamma between the first J and the second T (J_1T_2). Figure S2 shows mean gamma as a function of gamma source and JOL timing. A 4 × 2 two-way, gamma source by JOL timing, ANOVA showed that both main effects were significant, $F(3, 99) = 40.29$, $MSE = .07$, $\eta_p^2 = .55$, for gamma source; $F(1, 33) = 75.90$, $MSE = .06$, $\eta_p^2 = .75$, for JOL timing. The interaction also was significant, $F(3, 99) = 15.98$, $MSE = .08$, $\eta_p^2 = .34$. For the immediate JOLs, the significance pattern from Tukey's HSD test was $J_1T_2 \approx J_1T_1 < J_2T_2 < J_2T_1$, which is consistent with the findings of previous studies. Indeed, the immediate J_2T_1 gamma was so high that the delayed-JOL effect could not be found, $t(33) = 1.38$, $p = .18$. The corresponding outcome for the delayed JOLs was $J_1T_2 < J_2T_1 \approx J_2T_2 \approx J_1T_1$ in which the pattern was less clear because all the gamma values were high, except for J_1T_2, which on all grounds was expected to be low.

Figure Captions

Figure S1. Mean gamma as a function of JOL timing and type of practice. Each vertical hash mark depicts the standard error of the mean. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

Figure S2. Mean gamma as a function of gamma source and JOL timing in Condition SJT. Each vertical hash mark depicts the standard error of the mean. J = JOL rating; T = Test; JOLs = Judgments of learning.

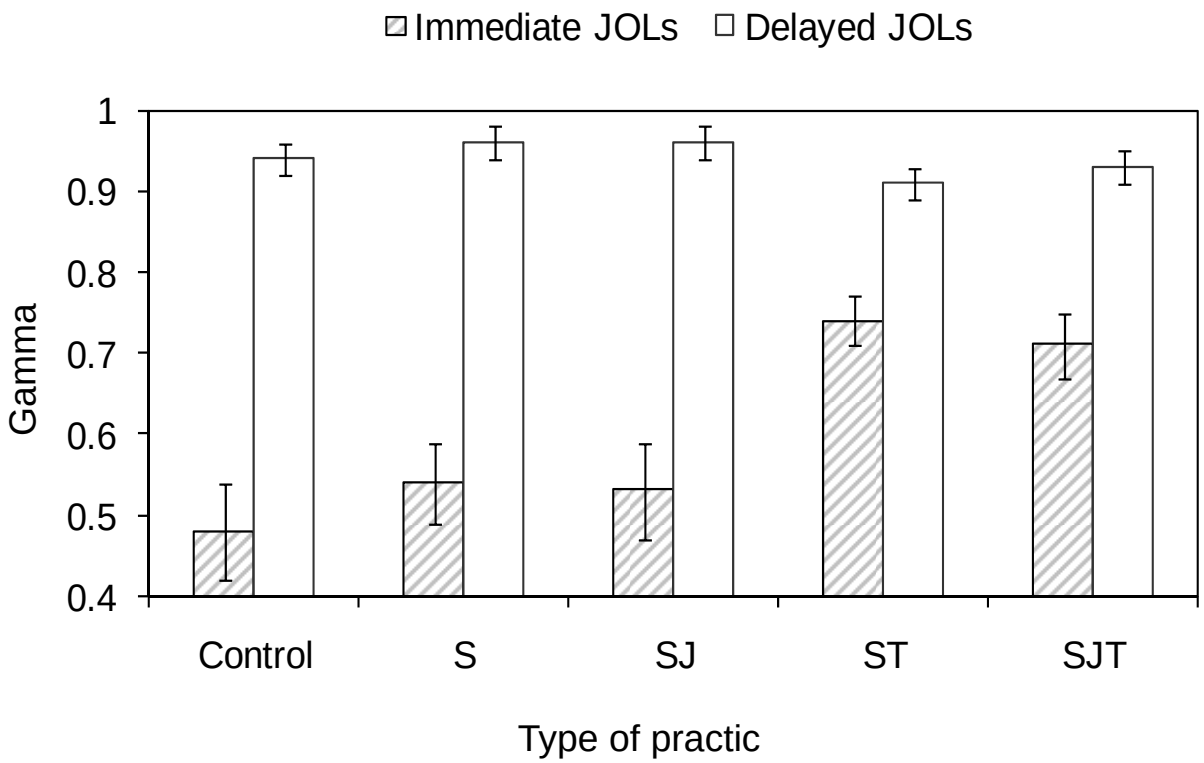
Immediate JOLs
Delayed JOLs
Gamma
1
0.9
0.8
0.7
0.6
0.5
0.4
Control
S
SJ
ST
SJT
Type of practic

Figure S1

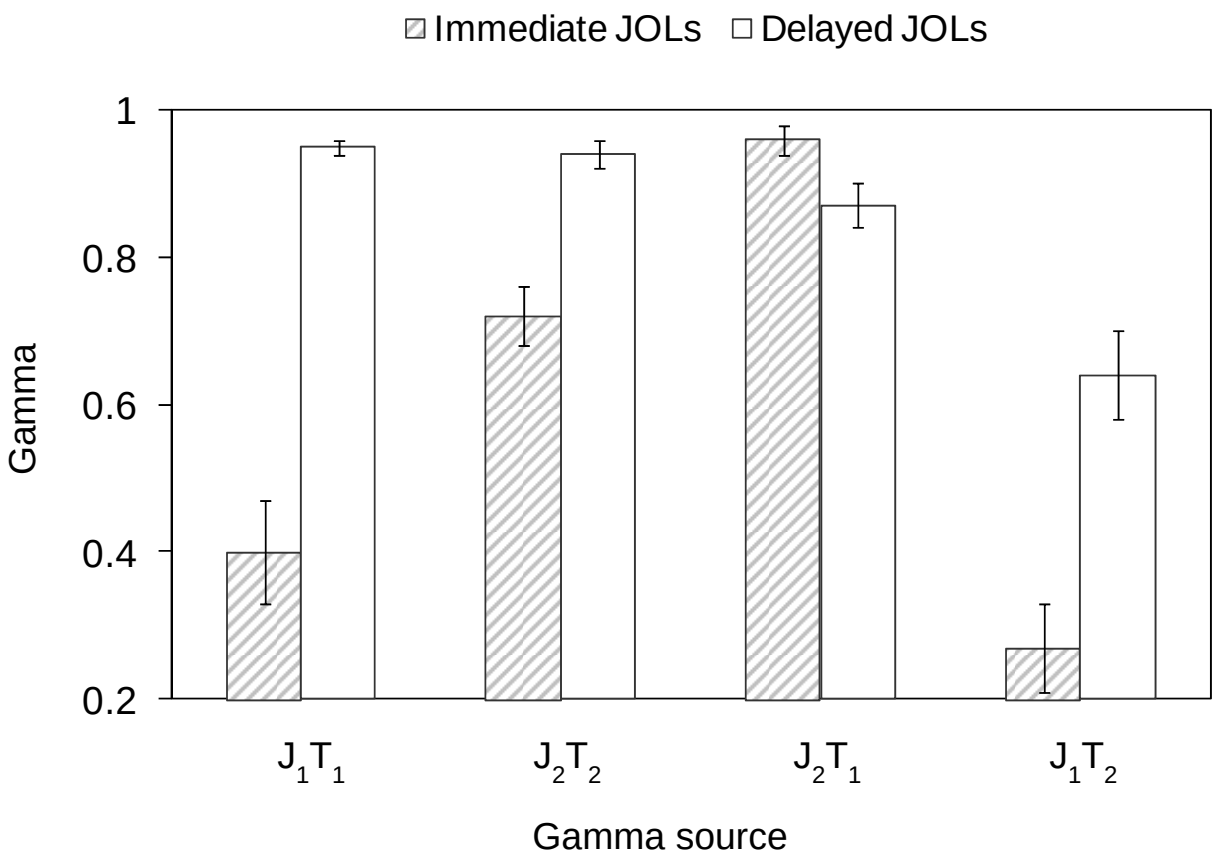
Immediate JOLs
Delayed JOLs
Gamma
1
0.8
0.6
0.4
0.2
J1T1
J2T2
J2T1
J1T2
Gamma source

Figure S2

Supplemental Material C (Jang, Wallsten, and Huber): Response Distributions in Conditions S, SJ, and ST, and the Predictions from the Best Fit Model for Conditions S and SJ

			Immediate JOLs						Delayed JOLs					
			0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
S	Empirical	Not recalled	.13	.19	.10	.03	.02	.02	.37	.06	.03	.01	.01	.01
	data	Recalled	.04	.10	.11	.10	.10	.06	.02	.04	.04	.06	.13	.22
	M4b (*confidence*)	Not recalled	.12	.17	.09	.04	.03	.01	.33	.09	.04	.01	.02	.01
	prediction	Recalled	.03	.11	.13	.10	.10	.07	.02	.05	.06	.04	.10	.23
SJ	Empirical	Not recalled	.12	.17	.10	.05	.02	.02	.32	.09	.04	.02	.01	.03
	data	Recalled	.02	.11	.14	.11	.07	.07	.01	.02	.02	.06	.15	.23
	M4b (*confidence*)	Not recalled	.10	.16	.09	.05	.03	.01	.29	.09	.05	.03	.02	.01
	prediction	Recalled	.03	.11	.13	.12	.09	.08	.01	.04	.06	.06	.07	.27
ST	Empirical	Not recalled	.11	.16	.06	.04	.01	.01	.23	.07	.02	.01	.01	.02
	data	Recalled	.03	.09	.11	.13	.09	.16	.02	.03	.03	.08	.12	.36

Note. M4b provides the best fit to the data in Conditions S and SJ as in the control condition. We do not report the model fits for Conditions ST and SJT because these conditions involved both the delayed-JOL effect and the testing-JOL effect, and we have no specific theoretical predictions for the combination of these effects.