Supplemental Online Materials

*Experiments 1-3 Similar Lure Rejections*

A "lure discrimination index" is typically calculated in this task (e.g., Yassa et al., 2011) by subtracting *p*("similar"|dissimilar foil) from *p*("similar"|similar lure), but this was unnecessary due to dissimilar foils having no association with a previous scene or delay, and therefore, the same value would have been subtracted from all similar lure rejection rates. Therefore, we simply analyzed *p*("similar"|similar lure), and these data can be found in Tables S1.

| | *p*("similar"\|target) | | | | *p*("similar"\|sim. lure) | | | |
|---|---|---|---|---|---|---|---|---|
| | Immediate | | Delayed | | Immediate | | Delayed | |
| | Reinst | Switch | Reinst | Switch | Reinst | Switch | Reinst | Switch |
| Experiment 1 | .13 (.04) | .16 (.04) | .21 (.04) | .24 (.04) | .44 (.06) | .49 (.07) | .34 (.06) | .37 (.06) |
| Experiment 2 | .15 (.05) | .12 (.05) | .19 (.05) | .24 (.07) | .38 (.06) | .43 (.07) | .30 (.07) | .29 (.06) |
| Experiment 3 | .21 (.05) | .26 (.06) | .27 (.05) | .32 (.05) | .53 (.06) | .52 (.06) | .44 (.05) | .41 (.06) |

Table S1. Erroneous "similar" rejections for targets and correct similar lure rejections (for similar lures) in Experiments 1-3 as a function of context reinstatement and delay. 95% CI in parentheses. reinst = reinstated, switch = switched.

In Experiment 1, similar lure rejections were less likely with context reinstatement ($F(1, 35) = 4.70$, $p = .037$, $\eta_G^2 = .01$) and after the delay ($F(1, 35) = 26.07$, $p < .001$, $\eta_G^2 = .08$) with no significant interaction ($F(1, 35) = .40$, $p > .250$). The reduced likelihood of rejecting similar lures might naturally follow from the increased false recognition of these items as "old", though this effect was not as robust as the effect of context on false recognition. In Experiment 2, context did not significantly modulate similar lure rejections ($F(1, 35) = .38$, $p > .250$), though these responses decreased across the delay ($F(1, 35) = 17.27$, $p < .001$, $\eta_G^2 = .08$), and there was no significant interaction ($F(1, 35) = 1.70$, $p = .201$). In Experiment 3, similar lure rejections decreased with the delay ($F(1, 35) = 21.82$, $p < .001$, $\eta_G^2 = .08$), but there was no significant effect of context reinstatement on similar lure rejections ($F(1, 35) = 1.34$, $p > .250$), and no significant context by delay interaction ($F(1, 35) = .32$, $p > .250$). High confidence similar lure rejections can be found in Table S2 Context did not affect high confidence similar lure rejections ($F(1, 35) = 1.35$, $p > .250$), but these were reduced with delay ($F(1, 35) = 41.72$, $p < .001$, $\eta_G^2 = .19$), and there was no significant interaction ($F(1, 35) = .64$, $p > .250$).

| | Immediate | | Delayed | |
|---|---|---|---|---|
| | Congruent | Incongruent | Congruent | Incongruent |
| *p*("similar"\|target) | .03 (.02) | .05 (.04) | .03 (.02) | .03 (.02) |
| *p*("similar"\|sim. lure) | .20 (.05) | .18 (.05) | .08 (.03) | .07 (.03) |

Table S2. High confidence erroneous "similar" rejections for targets and correct similar lure rejections (for similar lures) in Experiment 3 as a function of context reinstatement and delay. 95% CI in parentheses.

*Experiments 1-3 Dissimilar Foil False Alarm Rates*

In order to see if participants were more carefully monitoring their memories in Experiments 2 and 3 because of the nonexclusivity and warning manipulations, respectively, we ran a one-way ANOVA on dissimilar foil false alarm rates (Table S3) with experiment as a between-subjects factor. There was a main effect of experiment ($F(2, 105) = 3.216$, $p = .044$, $\eta_G^2 = .06$), explained by a significant difference between Experiments 1 and 2 (95% CI = [.00, .07], $t(70) = 2.03$, $p = .047$, $d = .48$) and a near significant difference between Experiments 1 and 3 (95% CI = [.00, .07], $t(70) = 1.89$, $p = .064$, $d = .44$). Therefore, our manipulations in Experiments 2 and 3 led to fewer errors to dissimilar foils than in Experiment 1, indicating that participants were more carefully monitoring their memories.

| | *p*("old"\|dissimilar foil) | *p*("similar"\|dissimilar foil) |
|---|---|---|
| Experiment 1 | .09 (.03) | .20 (.06) |
| Experiment 2 | .05 (.02) | .13 (.04) |
| Experiment 3 | .05 (.01) | .20 (.04) |

Table S3. False alarms and erroneous "similar" rejections for dissimilar foils in Experiments 1-3. 95% CI in parentheses.

*Experiment 2 Nonexclusive Target Results*

Data for nonexclusive targets can be found in Table S4. Recognition of nonexclusive targets (i.e., "old" responses) was modulated by reinstatement ($F(1, 35) = 8.64$, $p = .006$, $\eta_G^2 = .03$) but not delay ($F(1, 35) = 1.26$, $p = .269$), potentially due to the more durable memory created from exposure to two exemplars. The interaction was not significant ($F(1, 35) = .93$, $p = .343$). No main effects or interactions were significant for erroneous "similar" responses to nonexclusive targets (all $F$-values < 1 and $p$-values > .1). Therefore, context reinstatement also modulated memory when multiple exemplars were presented within a given context.

| | Immediate | | Delayed | |
|---|---|---|---|---|
| | Reinstated | Switched | Reinstated | Switched |
| *p*("old"\|nonex. target) | .81 (.06) | .71 (.08) | .76 (.07) | .70 (.07) |
| *p*("similar"\|nonex. target) | .15 (.05) | .20 (.06) | .17 (.06) | .16 (.05) |

Table S4. Responses to nonexclusive targets as a function of context condition and delay. 95% CI in parentheses. nonex. = nonexclusive.

*Experiment 1 vs. Experiment 2 False Alarms to Similar Lures*

In order to test if the nonexclusivity manipulation in Experiment 2 was able to reduce recall-to-reject, we ran a 2 (context: reinstated, switched) × 2 (delay: immediate, delayed) × 2 (experiment: Experiment 1, Experiment 2) ANOVA on similar lure false alarm rates (corrected with dissimilar foil false alarm rates) to compare results between

experiments. There was a main effect of context ($F(1, 70) = 39.60$, $p < .001$, $\eta_G^2 = .04$) and experiment ($F(1, 70) = 6.76$, $p = .011$, $\eta_G^2 = .08$) with no significant effect of delay ($F(1, 70) = 2.31$, $p = .133$). This main effect of experiment was explained by greater false alarms in Experiment 2 (see Figures 1a and 1b in main text). The context by delay ($F(1, 70) = .454$, $p > .250$), context by experiment ($F(1, 70) = 1.83$, $p = 1.804$), delay by experiment ($F(1, 70) = 1.41$, $p = .238$), and three-way interaction were all nonsignificant ($F(1, 70) = .04$, $p > .250$). Thus, the nonexclusivity manipulation was successful at reducing a recall-to-reject strategy, thereby increasing false alarms.

*Experiment 3 Warning Instructions for Memory Test*

Your memory will now be tested for the objects presented yesterday and today. You will be presented with three types of objects, again superimposed over a scene. Some objects will be similar (e.g., a cat posing differently or colored differently from the one in Phase 1 or Phase 2), some will be the same as studied (e.g., the same cat picture), and some will be new (e.g., a brand new object like a leprechaun). Your job is to press "similar," "old," or "new" for each object. Then you will rate your confidence (the options will be guessing, low confidence, medium confidence, high confidence, and certain). Similar and old objects might or might not be displayed with the original background that they appeared on in Phase 1 or Phase 2.

Please note: Research shows that a similar but never before seen object will often be mistakenly remembered as an old object if the background scene is the one originally associated with the old, related object (e.g., a black cat in outer space remembered as "old" because a grey cat had been originally paired with the outer space scene). We want you to try to avoid this mistake. Thus, if you remember an object-scene association but are unsure if the object is different or old, try not to be fooled into thinking that it was the same old object. Instead, keep in mind that sometimes a similar object will be presented, and other times an old object will be presented, and for each of these, the background scene might match the one in Phase 1 or Phase 2, or it might differ from the one in Phase 1 or Phase 2. Your job is to indicate whether

the object is similar, old, or new and to carefully evaluate the object and the scene to be as accurate as possible. This task will be self-paced and the response options will be displayed on the screen.

*Experiments 1-3 Response Latencies for Similar Lures and Dissimilar Foils*

A reviewer suggested that if participants were monitoring their memories for perceptual details, then response latencies should be greater for similar lures compared to dissimilar foils. Furthermore, Experiment 3 should have the greatest latencies because of the warning. We ran two 2 (item type: similar lure, dissimilar foil) × 3 (experiment: Experiment 1, Experiment 2, Experiment 3) ANOVAs on median response latencies including all responses to similar lures and dissimilar foils and only including correct responses (i.e., *p*(“similar”|similar lure) and *p*(“new”|dissimilar foil). These data can be seen in Figure S1.

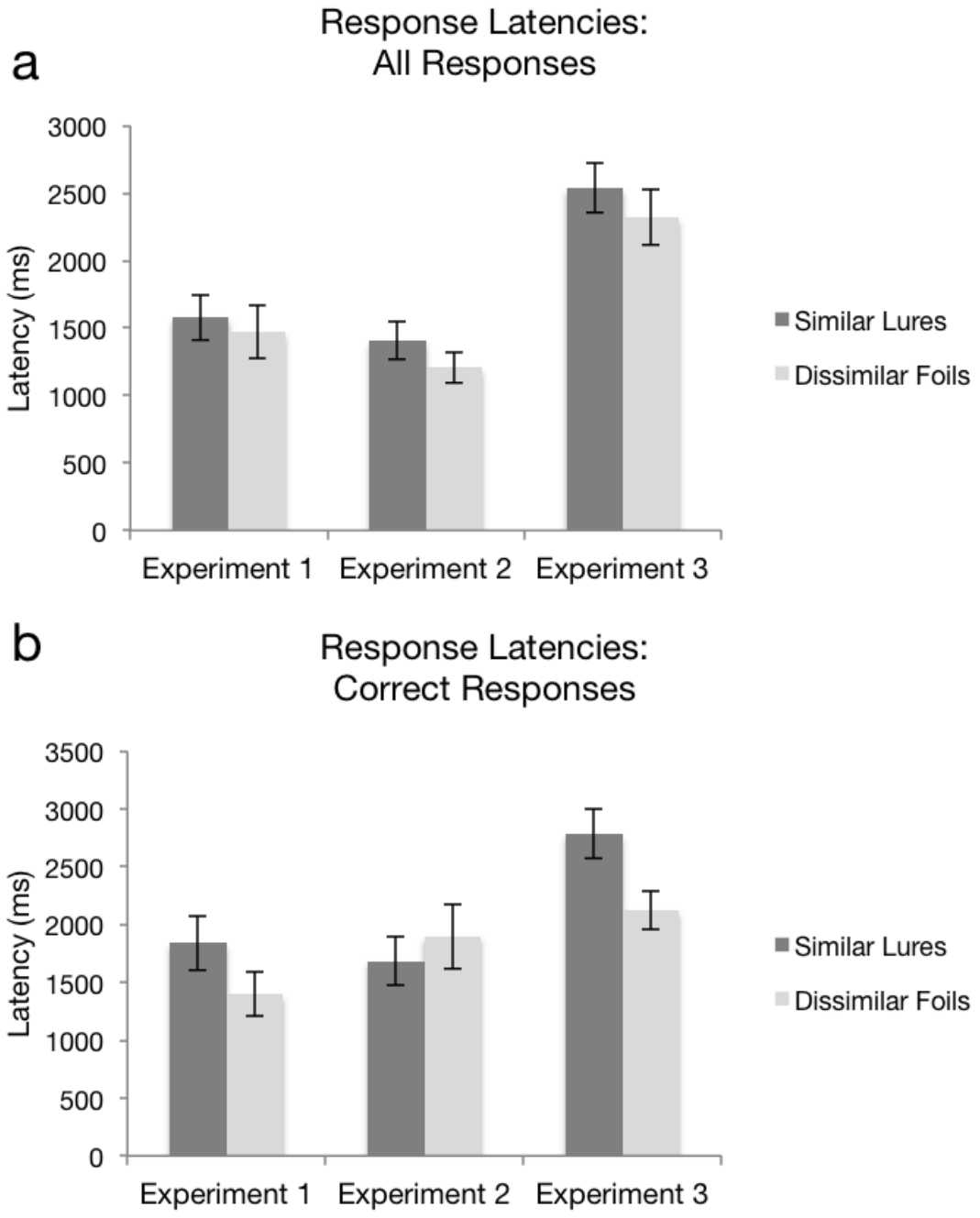



Figure S1. Response latencies to similar lures and dissimilar foils when (a) all responses are included and (b) when only correct responses are included.

One participant from Experiment 1 was excluded for not having any correct rejections of dissimilar foils, so Type III sum of squares was used for unbalanced groups. Both the ANOVA including all responses and the ANOVA including only correct responses revealed main effects of item type (all responses: $F(1,104) = 42.60$, $p < .001$, $\eta_G^2 = .03$; correct responses: $F(1,104) = 31.30$, $p < .001$, $\eta_G^2 = .05$) and experiment (all responses: $F(2,104) = 55.19$, $p < .001$, $\eta_G^2 = .49$; correct responses: $F(2,104) = 20.66$, $p < .001$, $\eta_G^2 = .25$). As can be seen in Figure S1, Experiment 3 had the greatest response latencies, suggesting that the warning made participants monitor their memories more closely. In all cases, similar lures had greater response latencies except in Experiment 2 when only correct responses were included. This is qualified by an item type by experiment interaction ($F(2,104) = 25.02$, $p < .001$, $\eta_G^2 = .08$), which was not significant when all responses were included ($F(2,104) = 1.46$, $p = .24$). Because Experiment 2 contained half the number of similar lures (in order to create nonexclusive target trials), the data from this experiment may have been less reliable. Furthermore, the differences between similar lures and dissimilar foils are quite robust in Experiments 1 (95% CI = [253.79, 627.13], $t(34) = 4.80$, $p < .001$, $d = .81$) and 3 (95% CI = [504.57, 816.13], $t(35) = 8.61$, $p < .001$, $d = 1.434$) compared to the reverse effect in Experiment 2 (95% CI = [4.66, 425.42], $t(35) = 2.08$, $p = .045$, $d = .35$). Thus, these data are highly indicative that participants monitored their memories for perceptual details when presented with similar lures.

*Pooled Experiments 1-3 Mnemonic Similarity Analysis*

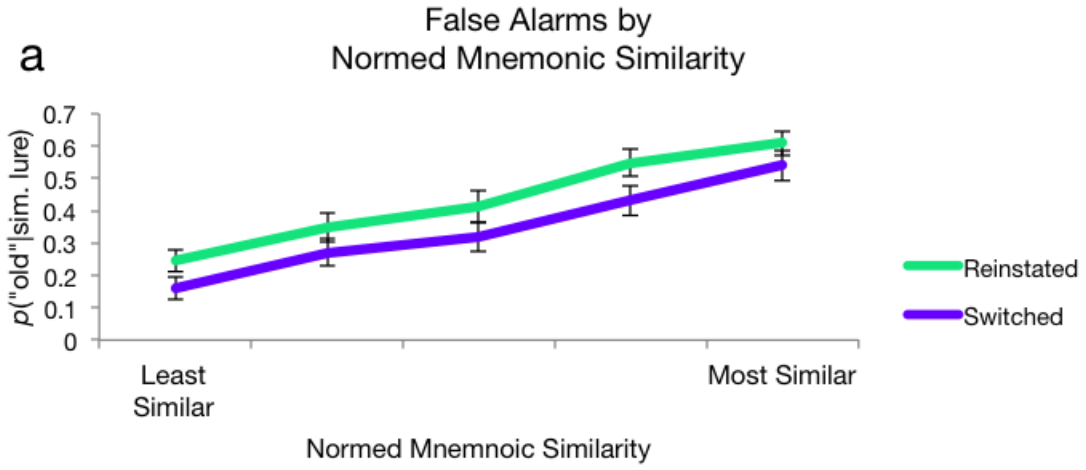
To further explore the context reinstatement effects in Experiments 1-3, we analyzed responses to similar lures as a function of the normed mnemonic similarity

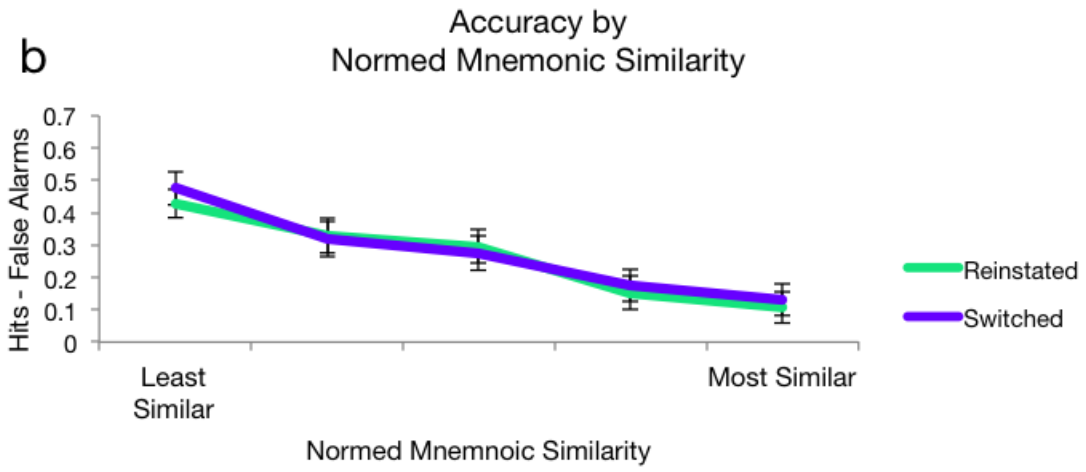
ratings obtained by Lacy et al. (2011). See Figure S2 for examples of object pairs at the five levels of similarity. These ratings represent the likelihood of confusing similar lures with studied objects in this task. Objects that are highly similar share perceptual and conceptual similarities, whereas those that are lower in similarity had greater perceptual and conceptual differences. For this analysis we pooled the data from Experiments 1-3, as this analysis could not be done in individual experiments due to the low number of items in each similarity level per experimental condition.


Figure S2. Examples of object pairs from the least similar (a) to the most similar (e). For more object pairs, see Craig Stark's website:

http://faculty.sites.uci.edu/starklab/mnemonic-similarity-task-mst/

We ran three post-hoc ANOVAs, one on similar lure false alarms ($p$("old"|similar lure)), another on accuracy ($p$("old"|target) - $p$("old"|similar lure)), and the last on similar lure rejections $p$("similar"|similar lure), including context, mnemonic similarity (treated as a continuous variable), and delay as factors. Mnemonic similarity did not interact with context in the analysis of false alarms ($F$(1, 107) = .03, $p$ > .250), and as can be seen in Figure S3a, false alarms were consistently elevated across all levels of similarity. Mnemonic similarity also did not interact with context for accuracy (Figure S3b; $F$(1, 107) = .15, $p$ > .250), but the analysis of similar lure rejections revealed a significant context by similarity interaction ($F$(1, 107) = 10.52, $p$ = .002, $\eta_G^2$ = .01). As can be seen in Figure S3c, context reinstatement was most likely to increase similar lure rejections for the least similar objects. That is, the similar lures that were least confusable with their targets also were the most likely to be affected by context reinstatement, leading to a correct "similar" response. This result suggests that the reinstated context could be used at retrieval to differentiate some of the most dissimilar objects, potentially based on item-context conceptual associations. For example, the sight of a ping-pong paddle and ball on the tundra scene might conjure up the thought that it would be easy to lose the ball in the snow. Upon seeing a contextually reinstated similar lure that was relatively low on the level of similarity, two paddles with no ball on the tundra, the original item-context association would be triggered, evoking a conceptual contrast between the two exemplars (i.e., noticing the missing ball). For the most similar objects (e.g., two harps positioned almost identically), it is less likely that such item-context associations could be made at encoding that would help differentiate the targets from lures at retrieval (Figure S2).


a
False Alarms by
Normed Mnemonic Similarity
p("old"|sim. lure)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Least
Similar
Most Similar
Normed Mnemnoic Similarity
Reinstated
Switched


b
Accuracy by
Normed Mnemonic Similarity
Hits - False Alarms
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Least
Similar
Most Similar
Normed Mnemnoic Similarity
Reinstated
Switched

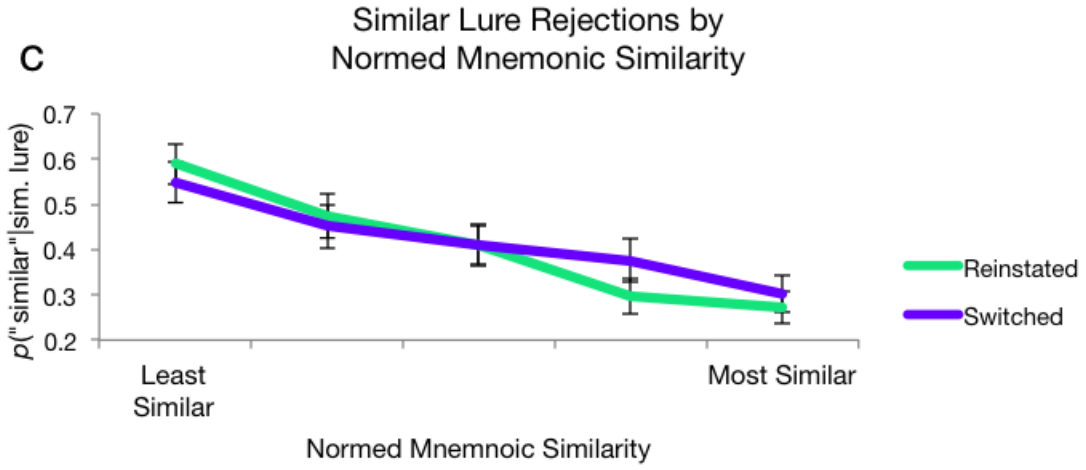

c
Similar Lure Rejections by
Normed Mnemonic Similarity
p("similar"|sim. lure)
0.7
0.6
0.5
0.4
0.3
0.2
Least
Similar
Most Similar
Normed Mnemnoic Similarity
Reinstated
Switched

Figure S3. Context reinstatement effects on (a) false alarms, (b) memory accuracy, and (c) similar lure rejections as a function of mnemonic similarity. Error bars are 95% CIs.

### *Experiment 4 Instructions for Object 2AFC Memory Test, Conceptual Familiarity, and Perceptual Similarity*

Your memory will now be tested for the objects presented yesterday and today. You will see two examples of each object, one that you have seen (e.g. a cat) and another that is similar but new (e.g. a cat posing differently or colored differently). Your job is to make three judgments. First, choose the one you had seen in Phase 1 or Phase 2 (Z = left option, M = right option). Note that the correct object can appear on either side of the screen, so the side of the screen should not affect your memory judgment. If you are unsure, please make your best guess.

Second, you will make a conceptual familiarity rating. Regardless of which of the two images were seen before, rate how familiar the object or concept depicted in the two images seems within this experimental context (using Q, W, E, R, T, Y; Q = strongly familiar, Y = weakly familiar). Familiarity is a feeling that you recognize this kind of object being presented before without necessarily recollecting specific details.

Finally, you will rate the perceptual similarity of the two objects (using 1, 2, 3, 4, 5, 6 at the top of the keyboard; 1 = highly similar, 6 = somewhat similar). Considering all pairs will have some perceptual similarity, please use the full range of responses. That is, do not simply use two or three buttons but rather use all six buttons.

Note that the object pairs might or might not be displayed with the same background that the original object appeared on in Phase 1 or Phase 2. Although this can sometimes help to cue your memory, all of the judgments should be made independent of whether the background was the same one seen with the original object from Phase 1 or Phase 2. This task will be self-paced and the response options will be displayed on the screen.

*Pooled Experiments 1-3 Context Reinstatement Effects Conditionalized on Item-Context Match*

A reviewer asked if the context reinstatement effects on hit and false alarm rates in Experiments 1-3 were larger for objects that were encoded on scenes that matched the object based on participant responses during encoding. That is, participants during the encoding phase were to determine whether or not an object belonged in each scene (yes/no; referred to as "matched" or "nonmatched). Across all experiments, there were more nonmatched than matched judgments (23.87% matched, 68.23% unmatched, and 7.90% non-responses because the encoding phase was timed). Furthermore, after conditionalizing each participant's data by their responses from the encoding phase, it was found that nine participants did not have any hit rates for at least one condition, and nine participants did not have any false alarm rates for at least one condition. Therefore, these subjects were excluded from the analysis. To analyze these data, we collapsed data across experiments and ran 2 (context: reinstated, switched) × 2 (delay: immediate, delayed) × 2 (item-context match: matched, unmatched) ANOVAs on hit rates and false alarm rates. These data can be seen in Figure S4.

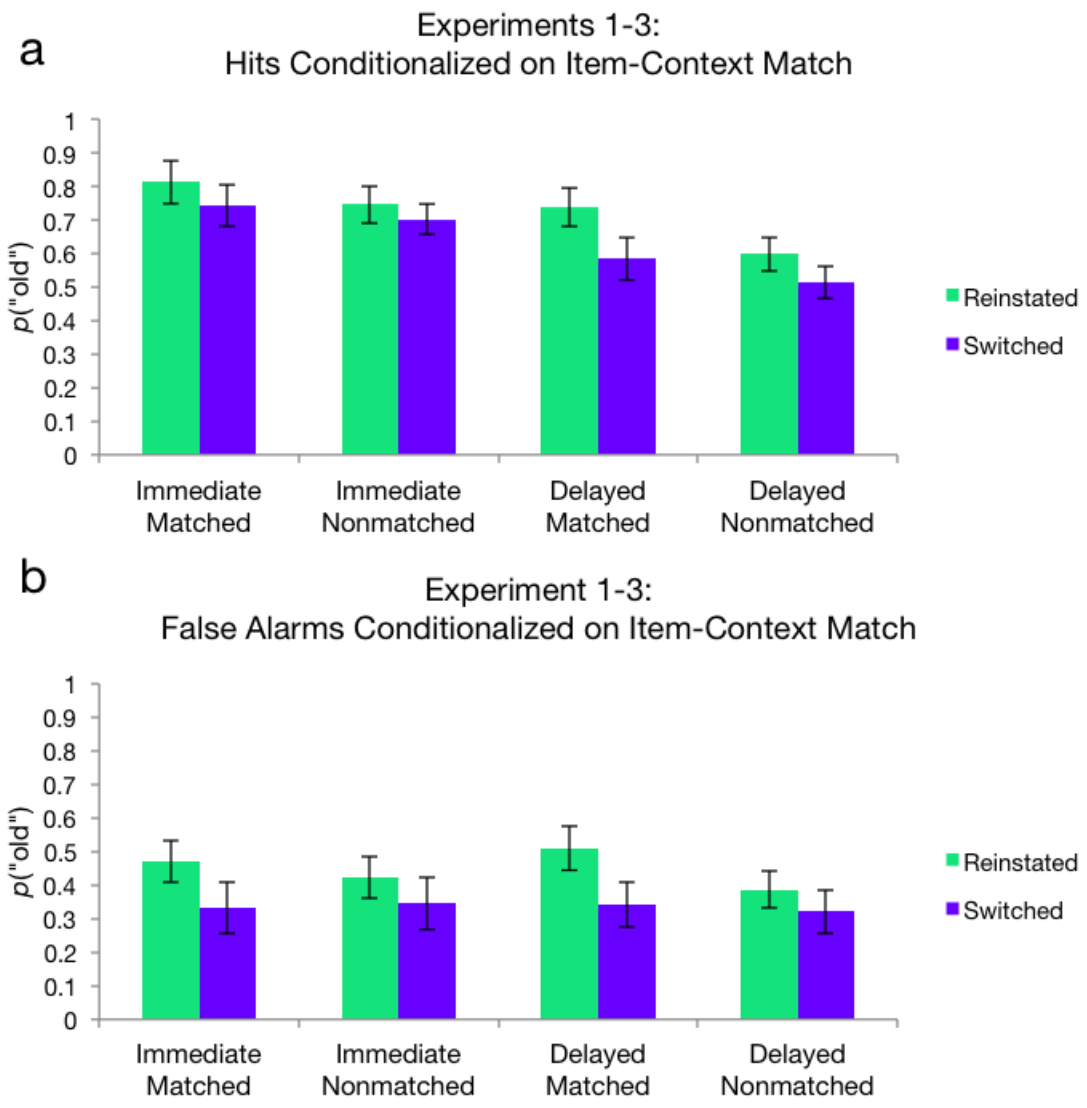



Figure S4. Context reinstatement effects become larger for (a) hit and (b) false alarm rates when conditionialized on whether the object matched its scene based on individual participant responses during the encoding phase.

Item-context match did indeed modulate the effects of context reinstatement. There was a main effect of context ($F(1, 98) = 33.21$, $p < .001$, $\eta_G^2 = .03$), delay ($F(1, 98) = 61.92$, $p < .001$, $\eta_G^2 = .08$), and item-context match ($F(1, 98) = 28.26$, $p < .001$, $\eta_G^2 =$

.03). The context by delay ($F(1, 98) = 4.75$, $p = .032$, $\eta_G^2 = .00$) and context by item-context match ($F(1, 98) = 4.143$, $p = .045$, $\eta_G^2 = .00$) interactions were significant, but the delay by item-context match ($F(1, 98) = 2.64$, $p = .107$) and three-way interaction ($F(1, 98) = .56$, $p > .250$) were not significant. The context by item-context match interaction appears to be from the fact that the context reinstatement effect is larger for matched compared to unmatched hits. Therefore, objects in an appropriate context are more likely to receive a boost in hit rates from context reinstatement.

The effects of context match on false alarm rates paralleled those observed on hit rates. There was a main effect of context ($F(1, 98) = 54.15$, $p < .001$, $\eta_G^2 = .05$) and item-context match ($F(1, 98) = 6.73$, $p = .011$, $\eta_G^2 = .01$) but not delay ($F(1, 98) = .027$, $p > .250$). The context by item-context match interaction ($F(1, 98) = 8.95$, $p = .004$, $\eta_G^2 = .01$) was significant, and there was a trend for the delay by item-context match interaction ($F(1, 98) = 3.04$, $p = .084$, $\eta_G^2 = .00$), but the context by delay ($F(1, 98) = .17$, $p > .250$) and three-way interaction ($F(1, 98) = .49$, $p = .484$) were not significant. As was the case for hit rates, the context by item-context match interaction suggests that the context reinstatement effect was larger for matched compared to unmatched false alarms. These findings again are consistent with the contextual distortion hypothesis, or the idea that context reinstatement can activate conceptual information about item-context bindings that drives false recognition.

*Pooled Experiments 1-3 Context by Delay ANOVAs*

The finding from Experiment 4 that object specific details degrade faster than item-context associations might suggest that the effects of context reinstatement should increase after a delay. Because there were trends for such an interaction (see Figure 1 in

main text), we pooled data from Experiments 1-3 and ran the context by delay ANOVAs on hit and false alarm rates. There was a main effect of context ($F(1, 107) = 34.88$, $p < .001$, $\eta_G^2 = .04$) and delay ($F(1, 107) = 100.67$, $p < .001$, $\eta_G^2 = .15$), and the interaction was significant ($F(1, 107) = 4.83$, $p = .030$, $\eta_G^2 = .00$). This interaction was qualified by a greater effect of context on delayed targets (95% CI = [.07, .14], $t(107) = 6.01$, $p < .001$, $d = .58$) compared to immediate targets (95% CI = [.02, .09], $t(107) = 2.90$, $p < .001$, $d = .28$). The ANOVA on the pooled false alarm rates revealed a significant effect of context ($F(1, 107) = 44.99$, $p < .001$, $\eta_G^2 = .06$) but not delay ($F(1, 107) = 2.462$, $p = .120$) or an interaction was not significant ($F(1, 107) = .344$, $p > .250$). Nevertheless, like hit rates, the effect size was larger for context effects on delayed items (95% CI = [.06, .13], $t(107) = 5.69$, $p < .001$, $d = .55$) compared to immediate items (95% CI = [.05, .12], $t(107) = 4.49$, $p < .001$, $d = .43$). Thus, because object specific details degrade faster, item-context associations are potentially used more for memory decisions after a delay.

*Data Coding*

For Experiment's 1-3 .mat files:

Column 1 = number assigned to object, Column 2 = description of object, Column 3 = image name of target, Column 4 = image name of similar lure, Column 5 = normed mnemonic similarity (1 = high similarity, 5 = low mnemonic similarity), Column 6 = image name of encoding scene, Column 7 = image name of scene when in switched condition, Column 8 = counterbalance list, Column 9 = can be a nonexclusive target in Experiment 2 (1 = yes, 2 = no), Column 10 = item condition (1 = target, 2 = similar lure, 3 = dissimilar foil, 4 = nonexclusive target in Experiment 2 only), Column 11 = context reinstatement condition (1 = reinstated, 2 = switched), Column 12 = delay condition (1 =

delayed, 2 = immediate), Column 13 = Experiment 1 or 2 identifier (of relevance to Experiments 1 and 2 only), Column 14 = memory response (f = old, g = similar, new = h for Experiments 1 and 2, g = old, f = similar, h = new for Experiment 3), Column 15 = memory response latency, Column 16 (Experiment 3 only) = confidence response (1! = low confidence, 5% = high confidence), Column 17 (Experiment 3 only) = confidence response latency.

For Experiment 4 .mat files:

Column 1 = number assigned to object, Column 2 = description of object, Column 3 = image name of target, Column 4 = image name of similar lure, Column 5 = normed mnemonic similarity (1 = high similarity, 5 = low mnemonic similarity), Column 6 = image name of encoding scene, Column 7 = image name of scene when in switched condition, Column 8 = counterbalance list, Column 9 = which side of screen (1 = left, 2 = right), Column 10 = context reinstatement condition (1 = reinstated, 2 = switched), Column 11 = delay condition (1 = delayed, 2 = immediate), Column 12 = which 2AFC test (1 = object 2AFC, 2 = scene 2AFC), Column 13 = memory response (left option = z, right option = m), Column 14 = memory response latency, Column 15 = conceptual familiarity rating (1! = high conceptual familiarity, 6^ = low conceptual familiarity), Column 16 conceptual familiarity response latency, Column 17 = perceptual similarity response (1! = high perceptual similarity, 6^ = low perceptual similarity), Column 18 = perceptual similarity response latency.

# References

Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T., & Stark, C. E. (2011). Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using high-resolution fMRI and variable mnemonic similarity. *Learning & Memory*, *18*, 15-18.

https://dx.doi.org/10.1101/lm.1971111