Dear Dr. Cushman,

We would like to thank the Reviewers for their detailed comments and supportive statements regarding our manuscript. We have incorporated their feedback in our revised manuscript, and have responded to each of their comments below. We strongly feel that the changes we have made in light of the reviewers’ comments have improved our submission. We are also happy to address any additional comments, or to clarify our responses if we did not address a concern to the Reviewer’s satisfaction

**Reviewer 1:**   
**Comment #1:** The authors state that among 48 participants they recruited, half of them   
were psychologically healthy, whereas the other half reported PTSD symptoms   
(p. 19). While the authors provide an assurance that the two groups of   
participants were very similar (p. 20), they do not provide details of the   
results, which makes it difficult to see how similar performance of the two   
groups were. In addition, the authors needed to exclude data from some   
participants, resulting in the total of 34 participants. With this   
relatively small number of participants, I wonder whether they had a   
sufficient power to detect any interactions between the group and their   
condition. More detailed descriptions about the group differences (e.g.,   
descriptive statistics for the key measures for each group) would be helpful   
to strengthen the paper. It would be also helpful if they could provide the   
number of participants in each group in the final sample.

**Response #1:** We now include information about group membership in the final samples on page 18. In the final sample of 34, 17 participants were from the PTSS group. Additionally, we include full reporting of our statistical analysis showing no effect of group in recognition memory, as well as in temporal context memory in the new Supplementary Results. We thought it best to include the details in in the supplementary materials, as we did not expect there to be major differences in episodic memory results between the healthy and PTSD-symptom groups.

**Comment # 2:** In analyses where they examined the effects of misattribution on   
recognition performance (e.g., p. 10), they used a fixed-effect analysis and   
used a bootstrap resampling procedure to test the generalizability. Given   
that the data have two levels (Level 1: trial; Level 2: participant), it   
seems that consideration of this multi-level structure is important during   
bootstrapping.

**Response #2:** Our motivation for the fixed effects analysis was that due to the nature of the predictors (3 alternative forced choice) and outcome variable (binary high confidence recognition memory), some participants had few, or even 0, responses in a particular cell (e.g. did not remember any CS- from pre-conditioning that were endorsed as being from post-conditioning). Resampling trials within each participant would not remedy the low number of trials in each cell, but by combining across participants there are enough trials in each cell for statistical analysis.

In the recognition memory and temporal context memory tests, each trial is an independent observation, such that what happened on the previous trial does not influence subsequent trials. Furthermore, we counterbalanced when each stimulus was encoded on day 1 and CS assignment across participants. The actual trial order of the stimuli (i.e. which stim was 1st) was randomized for all participants for all phases of the experiment. Thus there are already several experimental design features that equate observations across trials. We have also previously used this technique in other projects to overcome similar shortcomings (i.e. few trials of a certain type; Hennings et al., 2020 *Neuropsychologia*, Kim et al., 2012, *PNAS*).

**Comment #3**: Were animal/tool pictures used as ‘old’ or ‘new’ counterbalanced across   
participants?

**Response #3:** Each subject saw the same collection of images during encoding. Critically, the order of the stimuli during encoding was counterbalanced across subjects. The same set of stimuli were always used as novel lures. We include this detail in the Methods section on page 19: “The same collection of pictures were used for each subject during encoding, but the trial order was counterbalanced between subjects.”

**(Minor) Comment #4:** Page 6. In the first paragraph of Results, the authors briefly summarized their method but this summary was a little hard to understand the whole   
aspect of the procedures. For example, it is not clearly mentioned that   
participants completed the pre-conditioning phase before the fear   
conditioning phase (which is the very important aspect of their procedures).   
The fact that participants completed an extinction phase was also not really   
clear until I read Materials and Methods, which made it difficult to   
understand the reason why there were three options for the source memory   
test. This section should be expanded a little more to include the key   
aspects of the procedures.

**(Minor) Response #4:** We appreciate the Reviewer’s catch. This brief summary laying the groundwork for the Results section was short on details and is hard to follow without knowing more about the task design (which isn’t described until the methods section at the end of the paper). We have revised this paragraph on page 6.

**(Minor) Comment #5.** The authors did not report any results concerning physiological responses to CS+, while they reported the corresponding results in their earlier paper   
based on the same sample (Hennings et al., 2020). Were the same patterns   
replicated within this sub-sample?

**(Minor) Response #5:** For completeness we include the SCR and shock expectancy data as a Supplementary Results with this reduced sample reported in Hennings et al., 2020. The results are consistent, such that subjects acquire conditioned SCRs and shock expectancy to the CS+ versus CS- that diminishes over the course of extinction. We did not collect physiology during the memory test.

-----------------------------------------------------------------------------------------------------------------

**Reviewer #2**  
**Comment #1:** The authors report that attributing a stimulus to the conditioning   
phase is associated with better recognition memory accuracy, and discuss   
this finding extensively. However, there is an equally interesting finding   
in that attributing a stimulus to the post-conditioning phase is associated   
with significantly worse recognition memory accuracy. Why is this finding not discussed more?

**Response #1:** We thank the Reviewer for their enthusiasm for this finding, because we also think it’s very interesting. We debated how much to discuss this finding in the original submission. We now devote a paragraph to this finding in our revision on page 15.

**Comment #2:** In relation to this, I would like some more details on the final   
regression analysis. You state that source misallocation was used as binary   
regressor. What is the rationale for excluding correctly allocated trials?   
Moreover, if I understand the analysis correctly, you obtain one beta for   
misallocation? If this is the case, can you really conclude that ‘… source   
memory attributions to the conditioning context are the most predictive   
factor …’ (page 13, lines 6-8)? Judging by Fig. 4 (left-most panel), it   
seems like source memory attributions to the post-conditioning phase is also   
a predictive factor. I would predict betas with significant, but opposite,   
signs for misallocations to the conditioning and post-conditioning phases,   
respectively. Please conduct this analysis as I believe it is much more   
informative.

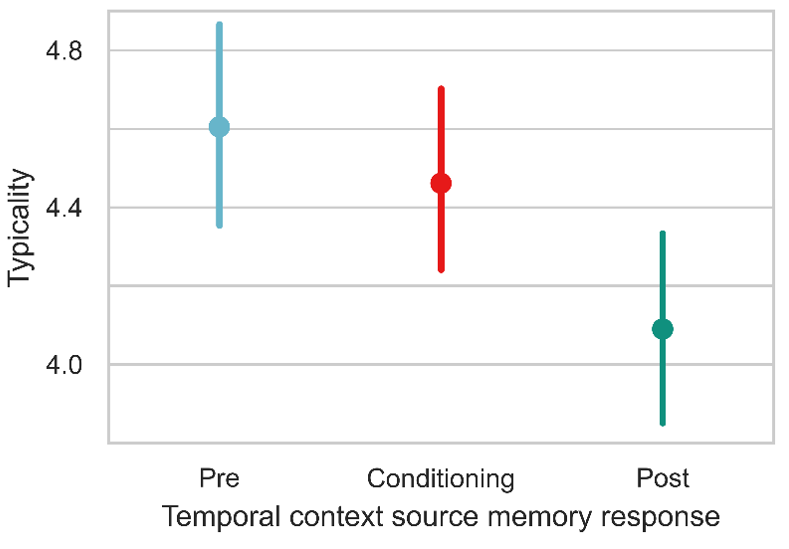
**Response #2:** We have run this analysis, adding misattributions to post-conditioning as an additional predictor in the model. As the Reviewer predicted, post-conditioning responses are a significant negative predictor of the retroactive enhancement (i.e. forgetting). When including this additional predictor, we found that the magnitude of the conditioning misattribution predictor was reduced, such that it was no longer stronger than CS type. We retained the analysis as the Reviewer suggests, as this analysis models various process impacting the retroactive memory enhancement effect. We have accordingly altered our interpretation of this finding in the discussion to focus on the finding that CS type and conditioning source misattributions have separable contributions. As a point of clarity, when the source memory responses are binarized correct trials (i.e. pre-conditioning responses) are not excluded, they are converted to a 0 response in our predictors of interest (all trials of pre-conditioning were entered into the analysis). We replicate the new figure below, and report the new statistics on pages 11-12.

****

**Fig 6. Source context misattributions to conditioning are as predictive of the retroactive recognition memory enhancement as conditioning itself.** All four factors were significantly predictive of recognition memory (above or below 0). Misattributions to the conditioning context was significantly greater than typicality. There was no difference between CS type and typicality. Misattributions to the post-conditioning context was significantly more negative than all other features. Black points and bars indicate the group mean and 95% CI.

**(Minor) comment #3**: What is the relationship between the source allocation and   
typicality? Are exemplars judged to be more typical also more likely to be   
attributed to the conditioning phase? Are less typical exemplars more   
attributed to the post-conditioning phase?

**(Minor) Response #3:** We thank the Reviewer for this suggested analysis. Their prediction is correct. We compared typicality ratings for items based on the temporal context source judgements. Interestingly, there was a significant stepwise trend across all phases, such that items that were sourced to pre-conditioning were judged to be more significantly more typical than items sourced to conditioning, and items sourced to conditioning were significantly more typical than items sourced to post-conditioning. We now include this exploratory analysis on pages 10-11 and in the new **Fig. 5C** (replicated below).



**C. Perceived typicality varied by temporal context source judgements**. Items that were sourced to pre-conditioning were also judged to be the most typical. The stepwise comparisons were all significant, such that pre-conditioning was greater than conditioning, which was greater than post-conditioning. Participant average data (with 95% CI) are shown, but statistics were run on trial-wise data in a mixed effects model.

**(Minor) Comment #4**: Could you please provide tables with the frequency of exemplars that fall within each category, e.g. on average (±SEM) how many stimuli of the   
conditioning phase were correctly recalled after being allocated to the   
post-conditioning phase?

**(Minor) Response #4:** We have added **Supplementary Table 6** which includes high confidence hit rate split by temporal context, CS type, and source memory response.

**(Minor) Comment #5:** It is a pity that no confidence judgements were included for the   
source memory allocations. To me, it seems like people have no clue where to   
allocate a stimulus (i.e. in Fig.2 all except one of the CS- allocations are   
no different from chance). I therefore wonder to what extent people actually   
believe they perceived a stimulus during the conditioning phase, and to what   
extent this is just a response bias. This should be discussed.

**(Minor) Response #5:** It is unfortunate that we didn’t gather confidence ratings in the source memory test. This is something we are now incorporating going forward. Overall, source memory is not very accurate. There is clearly a bias to endorse CS+ items as having been encoded during fear conditioning. Yet, the critical finding we dig into here is that the strength of this bias correlates with the strength of selective enhancements in memory for CS+ items encoded even before or after fear conditioning. As we note in our discussion, the causal relationship here is difficult to pin down. That is, do people remember CS+ items from before or after fear conditioning because the item was (falsely) attributed to the conditioning temporal context, or do people remember the item and think to themselves “I remember this, so I must have seen it while I was getting shocked”? We dichotomize this as a difference in a consolidation versus a retrieval based account (page 14). We’re not sure if confidence ratings would resolve this question, per se, but we understand that they would have been informative on some matters.

We include some more discussion on this in the revision on page 15, and note the addition of confidence judgements on source memory could shed more light.

**(Minor) Comment #6**: Why is there a mismatch between the false alarm rates in Supp. Table 1 and the proportion of DO responses for the novel lures in Supp. Table 2? Isn’t the FA rate defined as the proportion of DO responses for novel lures? Am I missing something?

**(Minor) Response #6:** We thank the reviewer for catching this. There was a small typo in the Python code that was used to create the tables, which did not impact any other analyses in the paper. The correction has been made to the table in the supplement.

**Stay healthy!**

--We’re trying!