Q1 Editor's Decision

We have four reviews back, which indicate reviewers are largely happy you did what you said you would. There are some points for revision; here are my further points and comments:

Remove exploratory conclusions from abstract. I am also a little uneasy about the amount of space they take in the discussion. When analytical flexibility is unleashed, and more complex analyses are used, it may be a relief that things make sense again. But we should go largely with the conclusions that follow from the pre-registered analyses themselves. I know your writing is already concise, and you may not like this, but can you halve the space given to the exploratory analyses in the discussion?

Response to Comment:

Thank you for your constructive comment. We have removed the exploratory analyses from the abstract and shorten the discussion regarding exploratory analyses in the discussion. Further, as suggested by Reviewer Prof. Dan Wright, we explicitly stated the failure of manipulation check in the abstract.

"Compared to the control group with no feedback, people who received feedback that indicated a tendency to make commission errors showed a shift toward a more conservative response criterion; in contrast, participants who received feedback indicating a tendency to make omission errors showed a shift toward a more liberal response criterion. However, manipulation checks failed to show that our manipulations had the expected effect on state memory distrust; we therefore did not find sufficient evidence that the effect of feedback was through its effect on state memory distrust toward commission or omission. Possible explanations for the results and future directions are discussed." (Abstract)

"***Evidence for Calibration based on Trait Memory Distrust Measures***

Trait memory distrust could be associated with metacognitive judgments through different mechanisms. First, trait memory distrust could be partly reflecting objective memory functioning. In this case, we would expect that for example, people with high distrust toward commission errors indeed are more likely to have more vivid recollective experiences and therefore higher recollection and belief ratings. Second, when motivated, people can also calibrate their belief judgments based on their tendency of committing memory errors, leading people with high distrust toward commission to down-regulate their belief ratings and people with high distrust toward omission to up-regulate their belief ratings. Our data suggest that people who considered themselves as more prone in general to making commission errors reported, on average, the highest recollection and belief ratings, while the reverse was true for those with higher distrust toward omission errors. Further analysis also showed that, when conditioned on recollection ratings, people with higher distrust toward commission errors reported on average lower belief ratings, whereas those with higher distrust toward omission errors reported on average higher belief ratings. Taken together, the results suggest that people indeed can calibrate their response criterion when being incentivized; however, the overall effect is much smaller than anticipated." (Page 34 Line 8-22)

"This study aimed to manipulate state memory distrust toward commission and omission separately, and examine their effect on criterion shift. Our findings indicate that feedback manipulation did influence response criteria: participants who were told they frequently made commission errors adopted a stricter response criterion, while those who received feedback about omission errors shifted toward a more liberal criterion. However, we did not find sufficient evidence that our manipulations had the expected effect on state memory distrust. Therefore, we could not offer evidence that the effect of feedback on criterion shift was through memory distrust. Overall, these results highlight the potential for feedback to influence metacognitive control of memory, but also point to the limitations in effectively manipulating aspects of state memory distrust separately. More research is needed to refine and improve state memory distrust manipulation and measurement." (Conclusion)

Q2 p 14 "After excluding three participants who had participated in similar studies and 28 participants who questioned the validity of the performance feedback, the final sample consisted of 622 participants" State that these exclusions were preregistered.

Response to Comment:

Thanks for this comment. We revised the sentence as follows:

"After excluding three participants who had participated in similar studies and 28 participants who questioned the validity of the performance feedback per our preregistration, the final sample consisted of 622 participants ($n_{\text{women}} = 306$, $n_{\text{men}} = 299$, $n_{\text{nonbinary}} = 2$, $n_{\text{other}} = 1$, $n_{\text{missing}} = 14$; $M_{\text{age}} = 38.67$, $SD_{\text{age}} = 12.23$)." (Page 12 Line 6-9)

Q3 The d' analyses were not pre-registered (as far as I can see) and should go in the exploratory analysis section.

Response to Comment:

Thank you for your constructive comment. The analysis on d' indeed was exploratory and only to serve as a check to show that the feedback mainly influenced response criterion. Given that this analysis was not preregistered, we have moved it to the supplementary materials (Page 45).

Q4 p 32 "This result suggests the effect of feedback on old-new recognition judgments was through its effect on metacognitive judgments of belief." There is no power calculation for this analysis so this conclusion should be deleted.

Response to Comment:

We have removed this statement from the manuscript.

Q5 Dan Wright suggests he should have asked for further manipulation testing if the manipulation check failed. IPA couldn't be given under these conditions because the Stage 1 should set in stone exactly what you will do next. You also did not say you would not do the remaining analyses if the manipulation check failed, so proceeding with the other analyses is right. However, you cannot interpret any effect of the manipulation as due to a manipulation of memory distrust. So, you have done the right thing. He also asks about other analyses you did but did not report; as these are not part of the Stage 1, it does not matter whether or not you did them. His suggestion that your abstract more closely reflect the Stage 1 I agree with, including reporting the failed manipulation check (and as I mentioned above, not the exploratory analysis).

Response to Comment:

Thank you for summarizing the reviewer's comments and sharing your thoughts. We agree that additional testing on the appropriateness and effectiveness of the manipulation check could have been the better approach, and would be more aware of this aspect in our future work. Regarding the suggestions on abstract, we now explicitly stated the failed manipulation check and removed the exploratory analyses (please see Response to Q1).

Q6 You don't need to do anything about the following, just a comment. Given what seemed what the clear conclusion would be before the data were in was that if the manipulation check failed according to your criteria, that is likely because the manipulation did not manipulate memory distrust, this is the most straightforward conclusion. The manipulation involved an easy to program deception. I sometimes wonder if participants believe anything a psychologist tells them given how often and freely, we try to deceive them. I suspect we pay the price for this: They treat us as seriously as we treat them. That makes it hard to answer one of Wright's main points: How to make a better manipulation. (Is it possible to do without deception?)

Response to Comment:

Thank you for sharing your thoughts on this matter. We agree that the most straightforward way to interpret the failed manipulation check is that the manipulation did not manipulate memory distrust, which we discussed in the discussion section. Regarding the use of deception in psychological research, we also agree that it can be that our participant pools might have grown suspicious of the instructions and feedback in our experiments and that research methodology avoiding deception should be preferred. Perhaps with regard to social influence on memory, we as community could look into the behavioral economics literature for inspiration.

best

Zoltan

Q7 Dan Wright

Note: I was a reviewer on the Stage 1 manuscript.

This is well written and the methods follow what I was expecting from the Stage 1 report.

I have two main comments. The first is of more importance.

1. The manipulations had little (if any) effect on people's responses to the manipulation check (or the memory distrust scale, but I'll focus here on the manipulation check). This was surprising, particularly for the manipulation check as those questions seem to be asking about the exact psychological constructs that the two non-control conditions are meant to affect. On lines about 411 of the Stage 1 manuscript it states:

Since the manipulation needs to reach a certain level of strength, only if the lower bound of the 80% CI on the effect size is above the minimal effect of interest (raw score difference of 1.6 with a SD of 2), will we consider the manipulation adequate. If the 80% CI is within the equivalence bounds [-1.6, 1.6], we will conclude that the manipulation did not reach an adequate strength.

A summary of the finding in the Stage 2 manuscript and the decision to still go ahead were:

We therefore concluded that the manipulation did not reach an adequate strength in manipulating state memory distrust. Despite the manipulation not producing the intended effects, we proceeded with the main analyses to examine potential effects that may still inform our research question.

I had to go back and re-read the relevant section of the Stage 1 manuscript. I had interpreted the phrase above as meaning if they found, according the criteria they state above, that the manipulation had not worked, they would work on making a manipulation that, according to their criteria, did work. I understand that having the data they went onto further analyses, but it seems this is now different. Of course, it may be that the manipulation checks and the memory scale do not work as they were expected to (e.g., people aren't aware of the influences of the manipulation), but that seems a question worth both addressing and trying to see if there is a manipulation check that does. The alternative is that the manipulation checks do work but the manipulation did not. If the manipulation had no (or little) effect, then the significant findings appear suspect. I went back to my original review to see if I raised issues with the manipulation check. I did:

"My concern at this point is whether the manipulation will have the desired effect on memory distrust that the authors believe. If I read their power analysis correctly, they believe it will almost completely account for memory distrust because they base their analysis on the memory distrust to response bias effect, if the causal chain above how they believe that this works. This makes the manipulation check critical.

As such, details of what counts for the manipulation check working is important for this. "

While the authors did state what they meant by failing the manipulation check, they did not examine, presumably through further data collection, why this occurred. I had assumed they would, but as a Stage 1 reviewer, I guess I should have asked them explicitly what they would do or requested a smaller manipulation check study before okaying the Stage 1 report. I am relatively new (as I guess we all are!) to reviewing these. My apologies for not requesting these. I had assumed failing the manipulation check would have prompted actions other than just continue, and that is my fault making that assumption. The editor may have views about what failing a manipulation check means for the rest of these analyses.

Response to Comment:

Thank you for your thoughtful comment. We agree with you on the importance of a valid and reliable manipulation check. Our overoptimism in the measures contributed to the current results which will -hopefully- improve our work in this domain in the future. Nonetheless, while the manipulation check results were not as expected, we maintain that our interpretation of the data remains valid. As the reviewers and the recommender have noted, the observed criterion shifts provided meaningful insights, even if the presumed mechanism—state memory distrust—was not confirmed.

Q8 On a minor note, in the stage 1 review they describe why it might be of interest to compare when the manipulation check was administered. In their code they list:

```
MC_Distrust_Commission <-
   afex::aov_car(Manipulation_Check_1~ Condition*Manipulation_order + Error(participantId),data = dat_S2)
```

This should be reported. I did find myself when reading the results going back to the code. The code page has a lot more analyses than appear reported.

Response to Comment:

Thank you for your comment. We did not report the said analyses in the main text given the results. However, we provided brief summaries in the footnotes (see Note 4 and 5). For both manipulation checks, we did not find support for a significant effect of manipulation order or an interaction effect with experimental condition. There are also other analyses conducted but not reported in the main text but in the supplementary materials. In the revision, we checked and removed any analyses not reported in the main text in the R markdown file and uploaded another R Markdown file that also include the analyses in the supplementary materials.

Q9 One other note. In the original footnote 3 (but not included in the new version for some reason) discusses some issues with the manipulation. Thus, the authors were aware that this could be an issue.

Response to Comment:

Thank you for your comment and your attention to detail. Previously in note 3, we briefly summarized the reasons why we decided on the final version of the manipulation (e.g., increase retention period and remove the example scenes in the feedback) given a previous round of piloting. The design and analyses were not part of the final experiment. We consider this was useful in Stage 1 for the reviewers to have a better idea of the design. In the stage 2 report, we removed it so not to confuse the readers with the non-central information.

Q10 If it is assumed that the manipulation worked (and therefore there was some unknown problem with the manipulation check), then the findings work as expected. I have two points. First, many of the outcome measures are dependent. For example, the c and beta. Similarly, when mixed models are done (and they aren't all linear, they are all generalized linear) statistics like d' and C are just coefficients in the probit regression (and really close in logit regression) if care is taken in coding the variables. Thus, these do not provide independent evidence and I assume the editor would prefer p-values only for one of these.

Response to Comment:

Thank you for sharing your expertise with us. We agree that the interdependence of tests should be controlled to properly control false positive rates. As stated in our manuscript, we have explicitly stated that our conclusions primarily relied on analyses of criterion c, rather than β.

“Given that β is a ratio statistic and more prone to bias, we draw conclusions primarily based on the results of criterion c.” (Page 26 Line 4-5)

In the revised manuscript, we also removed the analyses on *d*’ since these were not preregistered and are unnecessary to make sense of the data.

Q11 The second point is that the author notes that the scores on the first recognition test (before the manipulation) could be used to predict the scores during recognition 2, and then measuring the additional impact of the manipulations would likely be more powerful. I skimmed the code file, but this is a long file. Were these done? This seems a better use of additional predictors than, for example, predicting the old/new response from belief ratings and other variables.

Response to Comment:

Thank you for your comment. We did not perform the analyses suggested in the previous manuscript. It would indeed be good to run analyses controlling for the baseline in many situations. However, in the current setup, participants received biased feedback during the first recognition test, followed by summative feedback at the end of the first test. As a result, the overall performance in the first test was also under the influence of the experimental manipulations. We, therefore, decided to not include these analyses after consideration.

 ****

Q12 In summary, the authors appear to have done what they said that they were going to do, just I had assumed (wrongly) what would occur if they failed their manipulation check. The topic they purport to be examining has changed. Their original title was:

The effects of memory distrust toward commission and omission on recollection-belief correspondence and memory errors

But since their manipulation did not seem to affect distrust as expected, the new title focuses on the manipulation. This seems a largely substantial shift between a Stage 1 and Stage 2 document, but I am new to these registered reports. Maybe this is common to change what the focus is on.

With that, and recognizing I am part to blame for not asking the authors to be explicit in what they would do if the manipulation check failed, I recommend acceptance with three caveats. First, I think it is important, assuming that this is accepted, that this aspect is made clear to readers at the start (e.g., in the abstract say the manipulation check failed, and that possible reasons for this are discussed). Second, I believe the paper would be much stronger if the authors address this issue (likely with new data) as part of this paper prior to publishing this manuscript. Third, given their original intent (agreed upon from the title in the Stage 1 manuscript) was exploring the causal chain: manipulation -> distrust -> memory errors, they should stress that this was not shown.

Response to Comment:

Thank you for your comment. We have updated the title of the manuscript to better reflect the content of the analysis and the conclusions drawn. It is our belief that the correspondence to the contents of the stage 2 report is more important than consistency with the stage 1 title.

Regarding the emphasis and discussion of the results, we followed your suggestion to make it explicit that the manipulation checks failed and we could not find sufficient support for the proposed mechanism (please see response to Q1).

We agree that we could draw stronger conclusions with new data. However, we also believe that it would be perhaps more cost-beneficial to aim for a new and stronger experimental manipulation (ideally without the use of deception) on inducing aspects of state memory distrust, together with a better manipulation check. In this case, a separate project built on the results and discussions would be more appropriate.

Q13 Romuald Polczyk

The impression of the great competence of the Authors which I gained after reading Stage 1, only deepened after reading Stage 2. I do not see any weak points in the analyses (in fact, in several places they went beyond my own analytical competence).

The manipulation check did not show the effectiveness of the manipulation. Despite this, the Authors made the difficult decision to continue the analyses. I agree with this decision. Nonsignificant p-values do not imply nonexistence of the effect. It is therefore possible that in reality the manipulation worked, but it was not possible to show this in the check. Still, the effects of the manipulation can be seen in the main analyses. After all, the method used for the manipulation check - individual sentences assessed on a Likert scale - is rather weak, and much weaker than the tests used in the main analyses. The Authors commented in the Discussion that it may be that "state memory distrust measure did not adequately capture the change in state memory distrust levels" (p. 31, l. 4-5; also p. 32, l. 20ff., and 'Limitations'). It may be added that a single question is a less accurately measure than a questionnaire. This does not mean that other interpretations of the lack of an effect in the manipulation check are invalid. But the explanation may also be the low reliability of the measurement in the manipulation check. Currently, the Authors mainly discuss the issue of the validity of the manipulation check questions. It may also be worth mentioning their possible low reliability. After all, the p-values were not far from the significance level.

Minor points

I suggest not reporting confidence intervals as '.00'. Actually, if it was indeed exactly zero, it would mean that the effect is significant. Perhaps increase the decimal places to three in such cases. Writing '< .01' would be problematic unfortunately, as this does not exclude the the lower bound is negative, which was not the case.

Response to Comment:

Thank you for your recognition of our effort and comments regarding the reliability of single-item measures. In the revised manuscript, we have addressed the potential low reliability of the single-item manipulation check measure as suggested ("Finally, the use of single-item measure may have also suffered from low-reliability, allowing too much noise in the response." Page 33 Line 23-24).

Regarding the confidence interval of eta-squared, since eta-squared cannot be smaller than zero, a lower bound of 0.00 does not mean that the effect is statistically significant. Thus, there is no apparent contradiction between the p values and CIs reported. We believe that reporting the confidence intervals as they are is most transparent and consistent. While we understand the concern, we respectfully propose keeping the current format.

Q14 Anonymous Reviewer

The aim of this study was to investigate how the experimentally manipulated state of memory distrust, specifically toward commission and omission errors, influences shift in response criterion during a memory recognition task. The authors proposed that an increase in memory distrust toward commission would lead to a more conservative response criterion, whereas an increase in memory distrust toward omission would result in a more liberal response criterion.

The hypothesized effects of the manipulation on criterion shifts were confirmed. However, the authors noted that there was no evidence to suggest that these effects occurred through changes in state memory distrust toward commission or omission errors.

The study followed previously approved procedures and analyses, and the results are presented clearly and thoughtfully. In my opinion, this paper makes a valuable contribution to the misinformation literature and will likely inspire future research in this area.

I have two observations that might be helpful for future studies:

1. Design of the manipulation of memory distrust:
   It might be easier to observe significant effects of the manipulation by adopting a within-subject design, where participants respond to the same two statements both before and after the manipulation. This could enhance statistical power and increase sensitivity to detect subtle changes in state memory distrust.
2. Content of the feedback:
   Highlighting both the strengths and weaknesses of participants' memory likely enhances the

> credibility of the feedback. However, including information about both aspects of memory distrust (commission and omission errors) might introduce ambiguity, potentially diluting the effect of the manipulation and making it harder to isolate changes in state memory distrust specific to either commission or omission errors.

Response to Comment:

Thank you for your comments. We agree with both of your points. A within-subject design of the manipulation check would likely allow us to detect subtle changes in state memory distrust. In hindsight, perhaps the feedback should only focus on one aspect of memory distrust so to not distract participants from the message.

We have added these points in the discussions in the revised manuscript.

“A second possibility is that including both aspects of memory distrust in the summary feedback might have introduced ambiguity, which unintentionally influenced participants’ responses to the manipulation checks.” (Page 33 Line 9-11)

Furthermore, we recommend that future research could explore improved manipulation procedures:

“Based on these findings, we recommend future research to build on our work as well as those of others to further refine the manipulation procedures (e.g., measuring state memory distrust both pre and post manipulation) and establish better state memory distrust measures, such as multi-item scale with established reliability and validity.” (Page 35 Line 9-11)

Q15 Greg Neil

As this is a stage 2 review, I have reviewed this manuscript according to the stage 2 crtiera, and categorised my comments within each of the criteria requirements.

Summary:

This study looks at whether giving false feedback about commission or omission errors influences participants' tendency to respond either more, or less, liberally to a memory test. The intended manipulation of state mistrust did not appear to be effectives, although changes in the criterion were detected across feedback conditions anyway. Overall, the paper is well constructed and well written, and I have no concerns relating to the stage 2 criteria.

1) Can the data test the hypothesis, by passing the outcome neutral criteria?

Yes, the authors had previously shown a very thorough approach to their planning, and their approach to the data has been carried out, with all criteria being accounted for. The approach to excluding participants was well justified, and sampling sizes were achieved.

2) Are the introduction, rationale and hypotheses the same as the stage 1 manuscript?

Yes, I could see no substantial changes from the stage 1 manuscript.

3) Were the registered study procedures adhered to?

Yes, the stage 1 study procedures were adhered to.

4) Where there any significant deviations from the analytical approach set out in stage 1?

The initial plan was followed, but additional exploratory analyses were conducted to investigate the failure of the manipulation to change experimental condition state mistrust ratings when compared to control state mistrust ratings. Personally, I found the additional analyses to be well-justified and appropriate to investigating the potential mechanisms at work, so in terms of a stage 2 review, I have no changes to suggest.

5) Are the conclusions justified on the basis of the data collected and the analysis used?

I believe they are. I agree with the authors that, on the basis of their analyses, the issue here may be that the state mistrust ratings are either not measuring what the authors' wanted them to measure, or else the criterion shifts did not occur because of changes in explicit state mistrust. However, overall, I believe that this manuscript fulfills the stage 2 criteria.

Response to Comment:

Thank you for your careful assessment of the manuscript and the recognition of our effort.