Thank you for your revision. You have a great job of scientifically motivating predicted and smallest effects of scientific interest, which is the hardest part of the process!

Q1. The main thing is that the Stage 1 should not discuss exploratory analyses as that muddies the waters as to what was pre-registered with analytic flexibility removed and what was not. You can simply report the exploratory analyses in the Stage 2 in a non-pre-registered results section.

Response to comment:

We removed the exploratory analyses from the analysis plan and the discussion of the corresponding hypothesis from the introduction. These results of these exploratory analyses will only be discussed in the discussion section of the Stage 2 report.

Q2. You use inference by intervals for the first two rows of the design table, then hypothesis testing against no difference for the following. I have just published a piece arguing against mixing inferential practices in the same paper (https://osf.io/preprints/psyarxiv/2dtwv). As the different tests use different DVs, the issues I raise there don't apply strongly, so you may keep the mixed inferential strategy if on reflection you think it best. But it is at least inelegant: Why use inference by intervals in some cases and hypothesis testing against no difference in others? There is no inferential reason I can make out. You have all the figures (smallest effects/predicted effects) you need for using either method. For what it is worth, I think inference by intervals the superior of the two: Having defined a minimally interesting effect, one may, because of high power reject H0 with a smaller effect than the minimally interesting one. Yet the correct procedure with hypothesis testing is to reject H0. So high powered hypothesis testing can lead to embarrassing inferential contradictions which won't occur with inference by intervals. So just think about what you really want to do here.

Response to comment:

Thanks for this important comment. We have changed the primary tests to inference by intervals as suggested. In this revised version, we updated the power analyses for state memory distrust and criterion c using simulation-based method following Riesthuis (2024). This also allowed us to address your later comment (how likely to find the estimation falls within the null interval/ equivalence bounds).

"**Power Analyses.** We performed simulation-based power analysis for minimal effect testing and equivalence testing (see Riesthuis, 2024 for the tutorial) for the pairwise comparisons of response criterion and state memory distrust between the (a) control and omission conditions and (b) control and commission conditions. Specifically, the minimal effect testing calculated the percentage of simulations wherein the lower bound of an 80% Confidence Interval (CI) of the effect was greater than our SESOI (i.e., c_{diff} = 0.06, raw score difference of state memory distrust = 1.6). The equivalence testing calculated the percentage of simulations wherein the 80% CI fell within the two equivalence bounds (i.e., c: [-0.06, 0.06], state memory distrust: [-1.6, 1.6]). This analysis on criterion c showed that when the true effect size is c_{diff} = 0.15 (Cohen's *d* = 0.5), a group size of 100 participants will have 80% power to detect the minimal effect. When the true effect size on memory distrust is a raw score difference of 2 points (Cohen's *d* = 1.0), a group size of 210 participants will have 80% power to detect the minimal effect of 1.6. We therefore decide to set the minimum number of participants as 210 per group.

For the analyses on criterion c, a group of 210 participants will have 96% power to detect the minimal effect of c_{diff} = 0.06. With this group size, there is a 54% probability that the 80% CI will fall between the equivalence bounds [-0.06, 0.06] when the true effect is 0. For the analyses on state memory distrust, with the same group size and when the true effect is 0, the 80% CI will fall between the equivalence bounds [-1.6, 1.6] almost 100% of the time.

Sensitivity analysis with G*Power 3.1 (Faul et al, 2009) showed that a sample of 630 (210*3) would allow us to detect a slope of 0.015 criterion (c) units/Likert unit of state memory distrust (commission or omission) in a linear regression examining the association between state memory distrust and response criterion c (α = .05 and 1-β = .80; See appendices- Sensitivity Analysis protocol)." (Page 11 Line 217-237)

"Only if the lower bound of the 80% CI on the effect size is equal to or greater than Cohen's *d* = 0.20 (difference in c is 0.06, assuming an *SD* = 0.30) for the pairwise comparisons (distrust-commission vs. control; control vs. distrust-omission), will we consider the hypothesis supported." (Table 1)

"Linear regression with response criterion c (SD = 0.30) as the DV and either state memory distrust toward commission or omission (SD = 2.00) as IV. 90% CI will be calculated for the regression coefficients to compare against the SESOI." (Table 1)

When I said report your units I mean e.g. "regression slope = .017 c units/Likert unit of MC question". I know psychologists never do this but it makes things clearer. Same when reporting means.

Response to comment:

Thanks for this helpful comment. We revised the description as suggested in the manuscript.

When using inference by intervals, report the probabilities of both a) obtaining the 90% CI within the null interval if there is no effect; and b) the probability of obtaining a 90% CI outside the null interval if the predicted effect sizes on H1 is true. a) gives the severity of the test for the theory that H1 is true (i.e. how likely it is to get evidence against the theory if it were wrong). I typically do this with monte carlo simulations.

Response to comment:

Thanks for this important comment. In the revised manuscript, we reported obtaining the 80% CI within the null interval if there is no effect besides the minimal effect testing against SESOI. We opted for 80%CI (instead of 90% or higher) as a result of balancing resource constraint and error rate control.