Thank you for your revision: three of the reviewers are entirely happy you dealt with their points. Wright has some optional further points. I just request some tidying up of the Design Table, making each row contained in itself, so that it has, for example, its own power analysis. My points in detail follow.

1) First two rows of design table:

State in the table itself the DV in analysis - SSMQ for distrust for omission and MDS for commission.

The Table refers to a 90% CI. State in the table itself for what comparison: for omission (omission vs no feedback) and for commission (commission vs no feedback). Each could be as part of a 2X2 ANOVA.

Most informative would be to express each scale as an average over items so that the raw units are the same as the response scale subjects actually use, e.g. -4 to + 4. Then convert the Cohen's d (0.8) to these raw units so it is clear how much change in units of the original response scale is involved. The CI can then be in raw units.

"If there are interactions between manipulation check order and feedback, we will examine the effects of feedback on manipulation checks separately for the before and after the second test conditions, and decide if it is appropriate to report the two conditions together based on the directions of the effect in each condition." There is some analytic flexibility here. Be explicit about the precise conditions for each if-then decision. (It might be simplest to average over order and explore order later in exploratory analyses?)

Response to comment:

Thanks for this important comment. To clarify, the manipulation checks involve single-itemquestions answered on a 10-point scale. This is distinct from the SSMQ or the MDS, which measure trait memory distrust and are not used for the immediate manipulation checks in our study. In this revised report, we made minor changes to make this clear, including in Table 1. Further, we discussed expected raw score difference in both the main text and the design table. Finally, we took your advice and changed the planned analyses to one-way ANOVA with only experimental manipulation. The potential effect of manipulation check order will be reported as exploratory in the Stage 2 report.

“Two one-way ANOVAs, each with the experimental condition (Feedback: commission vs. omission vs. no feedback) as a between-participant factor will be run for the two manipulation checks (MC-commission and MC-omission) separately. The effect of manipulations will be reported averaging over manipulation check order conditions (also see Table 1).” (Page 18 Line 388-391)

“Assuming a SD of 2 scale points, this lower bound would translate to a raw score difference of 1.60 points on a 10-point scale.” (Page 18 Line 397-399)

“Then we will run one-way ANOVAs with the experimental condition (Feedback: commission vs. omission vs. no feedback) as a between-participant factor on β and c in the second test respectively. The effect of manipulations will be reported averaging over manipulation check order conditions.” (Page 18 Line 402-405)

Power/sensitivity needs to be determined for each row of the Design Table. That means for each confidence you need the power to detect the lower bound of the CI being above the minimal cut off. This entails two estimates: What is minimally interesting to define the lower cut off, plus an estimate of the likely effect - I am not sure if you can get this from your past studies?

Response to comment:

Thanks for this important comment. In the previous manuscript, we only defined the minimally interesting effect regarding the manipulation on response criterion shift. In the revised manuscript, we provided an estimation of the likely effect based on previous research using other methods to study criterion shift and then planned the sample size taking into consideration these two types of estimates.

The section is now expanded as follows:

"***Sample Size Justification***

Expected Effect Size of Criterion Shift. We expected the effect of the memory distrust manipulation on criterion shift to be smaller than the effect of explicit instructions to respond liberally or conservatively ($c_{a\ conservative}$ = 0.34, $c_{a\ liberal}$ = -0.50, $c_{a\ diff}$ = 0.84, Azimian-Faridani & Wilding, 2006) but larger than the effect of implicit biased feedback (Experiment 2: $c_{a\ conservative}$ = 0.39, $c_{a\ liberal}$ = 0.02, $c_{a\ diff}$ = 0.37; Experiment 3: $c_{a\ conservative}$ = 0.19, $c_{a\ liberal}$ = 0.05, $c_{a\ diff}$ = 0.14, Han & Dobbins, 2008). We therefore set a conservative expected effect size as a difference of c = 0.15 between the omission condition and the control condition and between the control and commission condition (in between Experiments 2 and 3 in Han & Dobbins, 2008).

Smallest Effect Size of Interest of Criterion Shift. We consider the following data pattern to be the smallest effect size of interest (SESOI) in the current experimental setup: the top 25% of participants who are most receptive to the memory distrust (Omission) manipulation will make one more hit response and one more false alarm response compared to their control counterparts in the current experiment. Similarly, the top 25% of participants who are most receptive to the memory distrust (Commission) manipulation will make one less hit response and one less false alarm response compared to their control counterparts[1]. The idea behind one hit and one false alarm comes from the idea that the remembrance of already one (false) detail can have practical value (e.g., misremembrance of the face of a culprit). Using the data from Study 2 of Zhang et al. (2023c) as the control condition and the expected differences between conditions, we created a synthetic dataset and calculated the SDT indices (See Table A1). In the synthetic dataset (See Table A1), the average difference of response criterion c is 0.06 between the Feedback-Omission and control or between control and the feedback commission with a standard deviation of 0.30 (i.e., Cohen's d = 0.06/0.30 = 0.2). With the same standard deviation assumed, our expected effect size would translate a difference of Cohen's d = 0.50.

We performed a priori power analysis for minimal effect testing (see Riesthuis, 2024 for the relationship between NHST and minimal effect testing) using G*Power 3.1 (Faul et al., 2009) for the pairwise comparisons of response criterion between the (a) control and omission conditions and (b) commission and control conditions. This analysis showed that a sample of 417 (139 in each condition) participants is required to detect the SESOI reliably (α = .05 and 1-β = .80; see appendix for the power analysis protocol).

---

[1] In reality, the probabilities will likely not be symmetrical. However, we consider the potential effect of the differences of probabilities on estimates insubstantial given that we set a rather conservative SESOI.

Expected Effect Size of State Memory Distrust. Given that in Dudek and Polczyk (2023), the difference of memory distrust between the experimental group and control group was close to that of Cohen's *d* = 1.0, we consider the expected effect size of manipulation on state memory distrust to be Cohen's *d* = 1.0.

Smallest Effect Size of Interest of State Memory Distrust. Assuming that memory distrust is the underlying mechanism of response criterion change, and that the correlation between state memory distrust and response criterion c is at least r = 0.25[2], this requires that the strength of the manipulation to be no smaller than Cohen's d = 0.2/ 0.25 = 0.8 for comparisons between memory distrust conditions and the control condition (calibration, Dienes, 2021). That is, state memory distrust toward omission should be 0.8 SD higher in the Feedback-Omission condition than in the control condition. Similarly, the state memory distrust toward commission should be 0.8 SD higher in the Feedback-Commission condition than in control condition. With a group sample size of 152 (456 in total) and a point estimate of Cohen's d = 1.0, the 90% CI will be [0.80, 1.20] with the lower bound being the SESOI (or, in terms of a raw point-difference on the original response scales, 90% CI [1.60, 2.40]). We therefore set the sample size to ***N*** = 456, the larger required sample size from the two analyses. Sensitivity analysis showed that a sample of 456 would allow us to detect a slope of 0.017 (α = .05 and 1-β = .80) in linear regression when examining the association between state memory distrust and response criterion c (See appendices- Sensitivity Analysis protocol)." (Page 9 Line 186 to Page 11 Line 232)

2) For your third row, specify how you will calculate the simple effects. As the conclusions actually follow from the simple effects, each simple effect could itself be the crucial test you jump straight to. This would give you more power. You then need a power calculation for the simple effect.

Response to comment:

Thanks for this comment. In the revised design table, we added the simple effects.

"One-way ANOVA (Feedback: commission vs. omission vs. no feedback) will be conducted for response criterion β and c.

Pairwise comparison would reveal that the response criterion is the most conservative in the feedback-commission condition, followed by the control condition. The feedback-omission condition will have the least conservative (most liberal) response criterion."

3) For the fourth row, can you state what raw regression slopes this corresponds to in meaningful units, so a plausibility judgment can be made about it really being minimally interesting?

Response to comment:

Thanks for this important comment. In the revised manuscript and design table, we added the information regarding regression slope based on estimated correlation and standard deviations.

[2] In Zhang et al., (2023), the trait memory distrust measure and memory task were measured three days apart and their correlation was *r* = .19. In the current study, the association is expected to be stronger given that we will measure state memory distrust right before or after the memory task and plan to increase participants' motivation to be accurate by raising the performance bonus to $3, 100% of the basic experiment payoff.

“We expect that the correlation between state memory distrust and response criterion c is at least r = 0.25 (raw slope = 0.0375).”

“Sensitivity analysis showed that a sample of 456 would allow us to detect a slope of 0.017 (α = .05 and 1-β = .80) in linear regression (See appendices- Sensitivity Analysis protocol)”

4) For the fifth, as a power analysis is difficult, consider as an exploratory analysis and remove from the table.

5) Same for the last row, state in raw units the change in regression slope you are powered to pick up.

Response to comment:

Thank you for your recommendations. Based on your advice, we removed the final two rows from the design table related to the analyses where conducting a precise power analysis proved challenging. These two analyses will later be reported as exploratory. Given that the hypothesis regarding recollection-belief association holds importance in the introduction, we decide to keep the discussion of this hypothesis in the stage 1 report but emphasize the exploratory nature due to the lack of adequate sample planning.

“**Exploratory Analysis**

The Effect of Feedback on Recollection-Belief Correspondence

“Even though we hypothesized about the effect of memory distrust on recollection-belief correspondence, the precise magnitude of this effect and the appropriate sample size required to test it remain uncertain. Given this uncertainty, we have decided to classify this analysis as exploratory.” (Page 19 Line 417-419)

It is unusual to be asked to present in raw units I know; but psychologists are so quick to remove all units from their analysis and thereby lose contact with their data as if the meaning of the units of their measurement scale made no difference whatsoever. As long as one cares about what size difference really matters, the units must matter.

by Zoltan Dienes, 27 Mar 2024 13:31
Manuscript: https://osf.io/4zecf
version: 1

Review by Dan Wright, 15 Mar 2024 15:59

Looking through mine and the other reviews, the authors have addressed most of the issues. My comments here are minor and/or suggestive. Most relate to the analysis.

1. Lines starting about 142. This was in an issue in the first version too. The authors bring up SDT. It is important that they relate this to a probit/logit regression since SDT (in the way they describe) is just a probit regression for each individual, and the flexibility of the later approach often makes it more useful (e.g., DeCarlo, https://www.columbia.edu/~ld208/psymeth98.pdf, Wright et al. https://pubmed.ncbi.nlm.nih.gov/19363166/).

Response to comment:

Thanks for this important comment. We truly appreciate the literature your provided. After consideration, we decided to keep the introduction as it is since the paragraph is to give the readers a general idea of the SDT approach, not the technical details.

2. Randomization is used in several places (e.g., when the manipulation check is used, it appears the order of the pictures is randomized and I think even which are used in which phase). This is just adding variation. Why is it being done? Further, how are the orders of these going to be taken into account in the models?

Response to comment:

Thanks for this comment. Indeed by randomization, we would add additional variation in the data. However, this is done to achieve stronger causal inferential power. That is, in this way, we could have more confidence to say that the observed pattern cannot be explained by the difference between test 1 and 2 or between the targets and fillers (since they are balanced between groups). Given our previous research, we do not expect the randomization procedure would have a substantial influence on the data (Zhang et al., 2024).

Reference:

Zhang, Y., Otgaar, H., Nash, R. A., & Rosar, L. (2024). Memory distrust shapes the dynamics of recollection and belief-in-occurrence over time. *Memory*. https://doi.org/10.1080/09658211.2024.2336166

3. Exclusions and sample size. The exclusions are listed, but it will be helpful to give an estimate for how many will likely be excluded and then use this to adjust the sample size. Also, you need to take into account the number of people who do not complete phase 2, and also increase the sample sought.

Response to comment:

Thanks for this comment. Given our previous experience with data collection on the Platform Connect, we estimate an attrition rate of 10% in the current study. We further anticipate most exclusions will be a result of failed attention checks in session 1. In the revised ms, we added the following information regarding valid responses from Session 1:

"Assuming an attrition rate of 10% from Session 1 to Session 2, we aim to collect 507 valid participant responses during Session 1." (Page 11 line 235-237)

4. I like the use of OASIS picture data base. This has variables on which the photos differ which likely relate to memorability. These should be included in the models.

Response to comment:

Photos from OASIS are accompanied with normative ratings of valence and arousal, which has been shown to influence memorability (e.g., Wakeland-Hart & Aly, 2023).

In our stimuli selection procedure, our analyses showed that on average the valence and arousal ratings in each block did not differ significantly. Further, in analyses on recollection and belief ratings, we have random intercepts for stimuli, which will take into account the differences among stimuli in e.g., belief judgments, which could have resulted from differences in valence and arousal. We believe that these safeguarding steps are sufficient for us to reach robust conclusions regarding our research questions.

We could further link the normative ratings from OASIS to the recognition memory data of the proposed study so that researchers who are interested in such question can make use of our data to answer their research questions.

Reference:

Wakeland-Hart, C., & Aly, M. (2023, September 12). Predicting image memorability from evoked feelings. https://doi.org/10.31234/osf.io/grxdz

5. The old/new, R, and K judgements will all be related. Given much of the introduction is about where these do not match (e.g., I remember clearly but know it did not occur), it should be described how the relationship among these will be examined. Also, it is important to discuss if there is an issue if there are not many of the discrepancies, since the introduction seems to be focused on discrepancies. I worry that the stats on the R and K judgements (after conditioning on Old/New and the other) become more about how people use the scale than their memories. Are there patterns of these that would lead the authors to stop the analyses?

Response to comment:

Thanks for this comment. We agree that the old-new judgment, recollection judgment, as well as the belief judgment will be correlated with one another. In fact, the differential association between recollection and belief in different experimental conditions are one of our hypothesis regarding the effect of memory distrust manipulation.

In our previous research (Zhang et al., 2024), we found that even in an experimental setting, in which the discrepancy between recollection and belief are rare (r = .96), the moderation effect of memory distrust on recollection-belief association was still observable. We therefore are confident that the current setup would allow us to detect an effect that is comparable to the one reported in Zhang et al. (2024).

Reference:

Zhang, Y., Otgaar, H., Nash, R. A., & Rosar, L. (2024). Memory distrust shapes the dynamics of recollection and belief-in-occurrence over time. *Memory*. https://doi.org/10.1080/09658211.2024.2336166

6. The authors include a random effect for pictures when analysing the R and K responses, but do not for the old/new response. Why? This contradiction requires explanation since it can be easily done predicting individual old/new responses. Also, as noted above there may be variables related to memory from OASIS.

Response to comment:

Thanks for this comment. If our understanding is correct, here you are referring to the analyses on response criterion indices, which will be based on the old-new judgments. In these two analyses, we will calculate beta and c for each participant using their Test 2 responses. Therefore, each participant will have 1 row of data only and we opted for ANOVA.

What you are suggesting, seems to be to change the method to multilevel model with recognition for each stimuli (Old vs. New) as the DV, the experimental manipulation as fixed effects, and participants as well as stimuli as fixed effects. This is a great suggestion, which we plan to run and report either as exploratory analyses or in supplementary materials.

For the point regarding the memorability of the photos, please see our reply to point 4.

7. It was not clear to me how the items on the two distrust scales will be used. I assume many of the individual items relate to items across the scales. Will these be combined and the psychometrics of the combined scale be reported. Obviously you would not treat them as distinct measures and analyse how they relate to the effects of interest independently.

Response to comment:

We will take the mean score of the items across each scale for subsequent analyses. A higher mean score would indicate higher level of memory distrust (either toward commission or omission).

We added this information in the revised manuscript:

"To ease the comparison of results, after establishing the internal consistency of the SSMQ and the MDS in the current sample, we will reverse-code the SSMQ then calculate the mean of all items for the two scales, so that a higher mean score in both scales reflects a higher level of memory distrust." (Page 13 line 282-285)