Dear Dr. Zoltan Dienes,

We appreciate your effort in securing high-quality reviews and your editorial decision. We have revised the Stage 1 report and wish to resubmit it to *PCI RR* for a second round of review.

In this revision, we updated the reasoning for sample-size planning, addressing the comments on manipulation strength and the smallest effect size. Furthermore, we revised the introduction section by providing the readers with more information regarding Blank (2009)'s theory as well as signal detection theory. We also added details of our manipulation and revised the exclusion criteria based on the reviewers' questions and advice. For some analyses, we found it difficult to offer sufficient justification for sample size planning and therefore decided to remove them from the Stage 1 report. Later, if these analyses were run, they will be reported as exploratory.

To increase participants' motivation to calibrate their recognition responses based on the feedback in the second test, we increased the performance payoff to $3, which should increase the association between state memory distrust and response criterion based on our hypotheses. Also, we added one question at the end of the experiment directly probing participants' motivation to calibrate their recognition responses to gain more information.

We appreciate your expertise in sample-size planning and theory-testing and would be happy to receive your advice on improving the current version of sample planning if you believe it is necessary before the execution of the study.

I look forward to hearing from you.

Best regards,

Yikang Zhang, also on behalf of Henry Otgaar, Robert A. Nash, and Chunlin Li

Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands.

Universiteitssingel 40, 6229 ER, Maastricht

E-mail : kang.y.zhang@outlook.com

Q1. Editor’s comments

First of all my sincere apologies for the long delay from receiving the referees' reports to me making a decision. On the bright side, I have 4 expert reveiws, and they are on balance very positive. In responding to the reviewers' points I want you also to take into account the following:

You should determine you have a satisfactory number of participants (e.g. power, if you are using power, as you are) for each row of the design table. That is you justify an effect size for each test, namely the test specified in each row of the design table, that tests the theory or hypothesis listed.

As power is the attempt to not miss out on interesting effects in the long run, you should specify the effect you just don't want to miss out on. Given a previous study found an effect of a certain size, given an effect a bit smaller would still be interesting, you should not simply use the effect from a past study. For ideas on how to approach this problem see: https://doi.org/10.1525/collabra.28202. It is often easier to think about this problem in raw effect sizes; Dan Wright illustrates by thinking about the actual numbers of words you are dealing with. Another example would be using raw regression slopes from your past studies so you can say what change in bias should follow from a unit change in distrust on the scales you use. Your past study may give you this. The bottom limit of the 80% CI on that slope may be (you would have to judge) roughly the smallest plausible effect size that may be still interesting. It would also tell you how much you would need to change distrust by in raw units to get a meaningful change in bias. And that might help answer the reviewer's question about how big the change in distrust should be to pass the manipulation check. (You can continue working with power and minimal effects; or you would switch to Bayes factors and expected effects.)

In sum, make sure you have considered the relevant effect for *each* test in the design table, so you have power calculations specificaly designed to address each test, in a scientifically justified way.

Response:

Thank you for your advice on sample size planning. In the revised manuscript, we planned the sample based on the smallest effect size of interest we wish to detect in the recognition data and defined the success criterion of manipulation given the expected association between memory distrust and response criterion. Furthermore, we elaborated why the planned sample size based on criterion shift would be sufficient for the other tests (e.g., recollection-belief association) in the analyses plan. For the GLMM analyses on commission and omission errors, we found it difficult to run power analysis confidently and could not provide sufficient evidence that the sample would be well-powered for these tests. Therefore, we removed these analyses from the Stage 1 report. These analyses, if done, will be reported as exploratory analyses in the Stage 2 report.

“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]. 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). Then we performed a priori power analysis using GPower 3.1 (Faul et al., 2009) for the main effect of feedback conditions on response criterion c showed that a sample of 456 (152 in each condition) participants is required to detect such an effect (α = .05 and 1-β = .90; see appendix for the power analysis protocol). Sensitivity analyses showed that a sample of 456 would allow us to detect an increase in R^2 in linear multiple regression no smaller than f^2 = 0.02-0.03 (α = .05 and 1-β = .90).

In the synthetic dataset (See Table A1), the average difference of response criterion c is 0.06 between Feedback-Omission and control or between control and feedback commission with a standard deviation of 0.30 (i.e., Cohen's *d* = 0.06/0.30 = 0.2). 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, the state memory distrust toward omission in the Feedback-Omission condition should be 0.8 SD higher than in control condition. Similarly, the state memory distrust toward commission in the Feedback-Commission condition should be 0.8 SD higher than in the control condition. Given that in Dudek and Polczyk (2023), the difference of memory distrust between the experimental group and control group was of similar magnitude, close to that of Cohen's *d* = 1.0, we consider the expected effect realistic in the current experiment setup." (Page 9 Line 187 to Page 10 line 211)

"Since the manipulation needs to reach a certain level of strength, only if the lower bound of the 90% CI is not lower than Cohen's *d* = 0.80[3], will we consider the manipulation adequate." (Page 18 Line 373-375)

"We did not perform a priori or sensitivity power analysis for this test given the complexity of the LMM model. However, results from Zhang et al. (2022b) suggest that the correlation between recollection and belief judgment show considerable difference between high (R^2 = 0.88) and low distrust people (R^2 = 0.95). We therefore consider a sample of 456 is well powered for this test." (Page 21 Table 1)

Table A1

Descriptives of the Synthetic Dataset

Condition	d'	c	β	n_{hit}	$n_{\text{false alarm}}$
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Commission	1.93 (0.69)	0.15 (0.31)	1.71 (1.37)	15.73 (2.61)	2.83 (2.42)

[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.

[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 raise the performance bonus to $3, 100% of the basic experiment payoff.

[3] With a group sample size of 152, the minimum effect that could be considered as adequate would be close to Cohen's *d* = 1.00, 90% CI [0.80, 1.20].

Control	1.87 (0.64)	0.08 (0.29)	1.30 (0.64)	15.94 (2.61)	3.15 (2.38)
Omission	1.86 (0.69)	0.03 (0.30)	1.18 (0.62)	16.14 (2.64)	3.49 (2.58)

Reviewer 1

Reviewed by Dan Wright, 31 Oct 2023 20:20

Q2. The submission describes the two main error types in memory recognition: false alarms (saying old to a new item) and misses (saying new to an old item). Lots of work has been done showing that it is harder to convince someone that something they remember did not occur, than to convince them that they do not remember has occurred (often explained by there being at least two types of memory). The proposed study uses false feedback during one recognition test stressing one of these two types of errors (misses and false alarms) to see how that affects the error patterns in a second recognition. The assumed causal chain is feedback -> a particular pattern of memory distrust (either focusing more on misses or false alarms) -> future memory recognition bias. There are lots of ways to change the response criterion in memory studies (e.g., telling them to do so), but this is examining if an indirect approach of feedback influences an asymetric type of memory distrust and this in turn has an impact on the memory recognition pattern. This is an interesting idea.

Understanding the feedback is therefore important to know if the manipulation will have a large enough influence on memory distrust. Here is what the document says: Feedback commission "they will receive false feedback on target items and true feedback on filler items. For each correctly recognized target, there is a 20% probability that participants will receive false feedback that this is actually a new scene. For each incorrectly recognized filler, participants will receive true feedback that this is a new scene." For example, suppose most people are about 75% accurate on hits and correct rejections. This would be 15 hits, so on 3 of these the person receives false feedback and presumably correct feedback on all 5 of the misses, but on all 5 of the false alarms they would receive correct feedback. Is this correct? So if they have been equal on the the hit and cr rates, then 8 of the 13 (62%) are of a commission error. The opposite would be true for the other group (with these assumptions). And these values will move around, but it is this percentage that presumably may affect the person belief in their asymmetry. If it only has a small effect then it would less likely to produce a detectable effect at the next stage.

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.

Response:

Thank you for your analyses.

In the design, we plan to only provide false feedback on hits and true feedback on false alarms in the feedback-commission condition. No correct feedback on misses will be provided. Similarly, participants in the feedback omission condition will only receive false feedback on correct rejection and true feedback on misses, with no feedback given for false alarms. So, we only provide one type of feedback for each condition. We further clarified this procedure in the revised manuscript:

“For participants who are in the feedback-commission condition, they will receive false feedback on correctly identified old items and true feedback on incorrectly identified new items. For each correctly recognized old item, there is a 20% probability that participants will receive false feedback that this is actually a new scene. For each incorrectly recognized new item, participants will receive true feedback that this is a new scene. No feedback on incorrectly identified old items (i.e., misses) will be given. For participants who are in the feedback-omission condition, they will receive true feedback on old items and false feedback on new items. For each correctly recognized new item, there is a 20% probability that participants will receive false feedback that this is actually an old scene. For each incorrectly recognized old item, participants will receive correct feedback that this is an old scene. No feedback on incorrectly identified new items (i.e., false alarms) will be given.” (Page 14 Line 292-302)

Q3. As such, details of what counts for the manipulation check working is important for this. The section on this in the proposal is too vague for me to know how this will be done. For example, they say "if the manipulation to increase distrust toward commission errrors is successful", so they need to define sucessful. They kind of define it in their power analysis. They use the r from Zhang et al (2023c) of .19 between memory distrust and response bias within this three group ANOVA design power calculation. If they are assuming the feedback effect is completely mediated by memory distrust (and if I read their lit review accurately this is at least the primary mechanism assumed), the "successful" manipulation check should require r values approaching one. This is unless I missed something. There is also the question of what to do if the manipulation is "successful" for one of the two manipulation check variables but the other only has, say, r=.7 (which I assume would be unsuccessful for them given their power analysis). It may be that the authors want to define "successful" at a lower level, but then they might not want to use the .19.

In summary, I think the effect of feedback may be smaller than what the authors believe it will be. Therefore I would want to see that it has the size effect that they are assuming before accepting the whole proposal. I focused on this aspect rather than on the lit review and other aspects of the methods because in an experiment clearly the manipulation is critical.

Response:

Thank you for this important comment. We have revised the sample size planning section in the revised manuscript. Please see our response to the editor above addressing your concern (Q1).

Q4. I did look some code for the simulated data out of interest. Consider the main results on recognition by condition. From what I could tell this is split by condition and/or isold, rather than (in lme4/R language)

m1 <- glmer(sayold ~ condition*isold + (1|item) + (1|person),family= binomial)

which allows this to be directly compared with the model without the interaction. Am I reading this correctly? Are the estimates of the response bias from the model

m1 <- glmer(sayold ~ isold + (1|item) + (1|person),family=binomial, data= just one condition)

the conditional modes for the person random variable, and then those being used in subsequent analyses?

Response:

Thank you for this important comment. In the previous manuscript, we did not hypothesize an interaction between condition and old vs. new. Therefore, we opted for analyzing hits/misses and false alarm/ correct rejection separately.

In the revised manuscript, we decided to remove these hypotheses from the Stage 1 report due to power considerations. If these hypotheses were to be tested, they will be reported in the Stage 2 report as exploratory and we will take your advice to run a complete model with two factors and their interaction first.

Q5. There is much work in memory using multilevel logistic models. I think this should be in the text of the proposal. Sifting through a webpage is difficult for me to follow without text explaining the steps, and I am a stats nerd. I wasn't going to go through this in detail since I think first I need to be convinced about the feedback having a big effect. It might also be useful to include two conditions where they are specifically told to raise of lower their response threshold, so that the size of the feedback effects can be compared with a more direct manipulation.

Response:

Thanks for this comment. We updated the script with annotations explaining the steps/ expected outcomes and will again update the scripts in the Stage 2 report to ensure the readability of the analyses.

Regarding the additional conditions with the explicit instruction, we agree that it would be interesting to compare the magnitudes of effects between different approaches. However, we are inclined to only focus on testing if the proposed manipulation of memory distrust would have an effect on criterion shift before moving on to the comparisons between different types of manipulations.

Reviewer 2

Reviewed by Romuald Polczyk, 17 Nov 2023 14:18

This project is about complex relationships of memory distrust with memory commission and omission errors and recollection beliefs.

I like the derivation of predictions - from theoretical considerations about autobiographical memory and autobiographical beliefs, to memory distrust and specific hypotheses. The hypotheses are based on both theoretical considerations and empirical data. I especially like that the hypotheses go beyond looking for relationships between variables but also include more complex moderation analyses.

In sum, I believe that the project is very well written, and I encourage acceptance of this proposal. The logic and plausibility of the hypotheses is excellent, as is the description of the planned procedure and statistical analyses. I especially appreciate giving access to the final Qualtrics procedure.

Below I include some remarks which perhaps may help to even improve the proposal:

Q6. I encourage the Authors to state more precisely how they will determine whether the manipulation check produced successful results - is a statistically significant effect enough, or should

the effect be of a certain size? For example, is a statistically significant change from an average score of about 2 to about 3 (on a scale of 1 to 10) enough?

Response:

Thank you for this important comment. Please see our response to the editor addressing your concern above.

Q7. I suggest including information about the reliability of the memory distrust scales.

Response:

Thanks for this comment. We have added the following information in the revised manuscript.

"Both the SSMQ and the MDS have been shown to have good internal consistency (e.g., Cronbach's α =.94 and .95 respectively, Zhang et al., 2023a) and criterion validity (e.g., recognition tests, Zhang et al., 2023b, compliance, Zhang et al., 2023c)." (Page 12 Line 258-261)

In the Stage 2 report, we will replace this information with the reliability indices calculated using the current sample.

Q8. p. 11: "After reading the information letter and giving informed consent, participants will first answer demographic questions about their age, gender, and education level, followed by the SSMQ and the MDS." - will the order of the SSMQ and MDS be randomized or fixed? Please specify.

Response:

Thanks. We added the information in the revised manuscript.

"participants will first answer demographic questions about their age, gender, and education level, followed by the SSMQ and the MDS in counterbalanced order" (Page 13 Line 266)

Minor points

Q9. p. 4: "For some events, people can hold strong beliefs about their occurrence without any recollections about them, such as the celebration of your first birthday (believed-but-not-remembered events)." - I doubt this relates to autobiographical memory at all; this is just knowledge about some past events, not remembering them. Like, say, knowledge about appendicectomy surgery - we may know that it took place, but we cannot remember it because we were under anesthesia.

Response:

Thank you for this comment. We agree that believed-but-not-remembered events are not autobiographical memories. In the revised manuscript, we revised the sentences to avoid this confusion.

"In most instances of remembering, the rememberer trusts that the recollected event truly happened in the past. Such event representations, which encompass both vivid recollection and a firm belief in the event's occurrence, are referred to as believed memories. However, there are other types of event representations in addition to believed memories. For some events, people can hold strong beliefs about their occurrence without any recollections about them, such as the celebration of your first birthday (believed-but-not-remembered events)." (Page 4 Line 47-52)

Q10. I extremely appreciate providing access to the Qualtrics procedure for reviewers. I just suggest enabling to proceed without giving answers on the questionnaires as it takes time.

Response:

Thanks for this tip. We will adjust the Qualtrics for reviewing in the future. To ease the access of the survey information, we uploaded word files of the survey on OSF (https://osf.io/su3vf; https://osf.io/kb627).

Q11. I could not locate any information about funding for the research. Although I don't think this is important and I have no doubts that the Authors will find funds, I mention this as reviewers are required to do so by PCI.

Response:

Thanks. We confirm that we have funding for the proposed research.

Q12. Although I by no means believe that every scientific project needs to be 'applied' to some degree, I encourage the Authors to think about possible applications of the results; or, to state clearly that they expect none of such.

Response:

We appreciate your suggestion. As also mentioned by Reviewer 3, memory reporting errors and criterion shift could be of interest to eyewitness identification and testimony validity assessment. We will keep this in mind and discuss the practical implications of the proposed research in relation to the legal arena in the Stage 2 report.

Reviewer 3

The proposed research concerns the effect of manipulating two types of memory distrust (i.e. concern about making omission errors or commission errors) on memory reporting (recognition, recollection, and belief judgments). The research problem formulated on the basis of the previous research results of the first author and their colleagues is very interesting. The findings of the study will expand our knowledge on the psychology of memory and may also have implications for the psychology of eyewitness testimony. The question and hypotheses were clearly presented and justified. The manuscript is concise and well-structured. The research procedure is presented in detail, and surveys from the Qualtrics research platform are included. Extremely careful and detailed data analysis was planned for each research question and interpretation of possible results was indicated.

Meeting the Stage 1 RR criteria:

1A. The scientific validity of the research question(s) – Yes. The research question aligns coherently with previous research findings and Blank's (2017) theory of recollection–belief divergences and validation.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable. – Yes. The research questions and the hypotheses logically derive from the theoretical introduction presented.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable). – Yes. The planned analyses will answer the questions posed. The sample size is sufficient to ensure a high probability that a significant result is true (power: 90%). Data exclusion criteria are specified, and the smallest effect size of interest is defined based on previous research.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses. – Yes. The research procedure is thoroughly described and made available. The analysis plan includes verification of all research hypotheses and state precisely which outcomes will confirm predictions.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s). – Yes. The authors planned to test the effectiveness of the experimental manipulation, and this manipulation was also tested in a pilot study.

I only have a few minor comments:

Q13. p. 7, l. 20 – 22: It is stated that the trait of memory distrust is related to the objective functioning of memory. However, it may be worth considering a slight modification to this statement, as certain studies have not demonstrated a clear association, for example, Kuczek, Szpitalak, & Polczyk, 2018.

Response:

Thanks. We have added the reference in the revised manuscript.

“First, given that trait memory distrust is associated with objective memory functioning, we can expect people who are high (vs. low) on memory distrust to perform worse in memory tasks (i.e., lower sensitivity, Zhang et al., 2023b; but see Kuczek et al., 2018 for a different result).” (Page 8 Line 153-154)

Q14. p. 15, Outliers and Exclusions section: It might be worth considering the exclusion of participants who guessed the research hypotheses, given that the description of the procedure suggests that participants will be asked about this matter. Consider providing additional clarification regarding what would be deemed a correct guess regarding the study's purpose, for instance.

Response:

Thanks for this important comment. In the revised manuscript, we added another exclusion criterion based on the open question regarding the purpose of the experiment. If participants correctly guessed the false nature of the feedback and did not believe the feedback, their responses will be excluded.

“Participants who indicate, in the open question regarding the goal of the experiment, that they distrusted the feedback they received will be excluded from the analyses. Two independent coders (Y. Zhang and one research assistant) will code participants’ responses regarding this exclusion criterion.” (Page 17 Line 357-360)

Q15. In relation to SDT in the context of response bias, it could be beneficial to mention what a c indicator is, similar to the explanation provided for the β indicator on page 6.

Response:

Thanks for this important comment. We have added the missing information in the data analyses overview section.

"Response criterion indices β and c will be calculated using the psycho package (Makowski, 2018), β is calculated based on the likelihood ratio of the two distributions (noise and signal) while c represents the distance between the response criterion and the unbiased point, expressed in units of standard deviations. A higher value of either β or c would indicate a greater tendency to recognize stimuli as new instead of old (i.e., a more conservative response criterion). Since β is based on a ratio and more likely to violate distribution assumptions than c (Zhang et al., 2023b), when the results of the two indices diverge, we will give more weight to c when reaching conclusions." (Page 16 Line 342-348)

Q16. There appears to be an inconsistency in the citation of Nash et al. between the text (2023) and the referenced literature list (2022).

Response:

We corrected the reference in the revised manuscript.

Reviewer 4

Dr Greg J Neil

This Stage 1 submission sets out a plan for a study to examine whether feedback about commission errors (remembering something that did not happen) and omission errors (forgetting something that actually happened) will influence participant's criterion placement, and also the correspondence between their beliefs about their performance and their recollection. I write this review from the perspective of someone who is familiar with the methods used in this paper, including signal detection, and the analyses proposed.

Overall, the logic and methods seem appropriate to the question. The scientific validity of the question is justified, the logic of the hypotheses tracks, and the analysis is appropriate. I do, however, have some comments on the clarity, and on the exclusion criteria and outcome-neutral conditions, which I set out below:

Q17. Clarity – This is a complex area, and consequently the introduction needs to give more explanation of were some key ideas required for a reader to understand the study. I believe this could be addressed by focusing on two parts of the introduction. First, the 2nd paragraph on p.4 introduces some key ideas which could be fleshed out and given some additional examples, to help readers to follow the central ideas of the paper.

Response:

Thank you for this important comment. In the revised manuscript, we changed the flow of this paragraph and elaborated on the Blank (2009) paper, which proposes a theoretical framework of remembering and belief formation in particular. The section now reads as follows:

"The dissociation between recollection and belief can also be observed in other ways. Scoboria et al. (2014) reported that memory characteristics that predict recollection ratings well (e.g., perceptual, re-experiencing, emotion intensity, event specificity) did not predict belief ratings and vice versa (e.g., plausibility judgment). Furthermore, studies on nonbelieved memory showed that autobiographical beliefs can be altered relatively easily —more so than can recollections—in response to social information contradicting one's memories (Li et al., 2020; Otgaar et al., 2013, 2017; Scoboria et al., 2018; Wang et al., 2017). In particular, the credibility of social information influences the formation of false autobiographical beliefs but not the recollective features (Scoboria et al., 2014). On the other hand, in the course of forming false memories, false beliefs can be created more easily relative to false recollections (Mazzoni et al., 2001; Pezdek et al., 1997). People who start to falsely believe a suggested event may then utilize general scripts for similar events, or memory details of other episodes, to develop their recollections (Pezdek et al., 1997).

Blank (2017) posits that these recollection-belief divergences, such as nonbelieved memories or believed-not-remembered events, arise from normal, healthy metacognitive monitoring and control processes that balance recollections and reality constraints. Specifically, Blank proposed a theoretical framework of remembering, in which, he distinguishes memory, belief, and the communication of memory (e.g., memory reporting) and elaborates on the role of social influence in remembering (2009). First, memory traces are activated through either internal processes (e.g., intentional searches) or external cues (e.g., verbal prompts). In the second stage, various information (e.g., physical evidence or social feedback) joins memory traces in the validation process, in which, the validity of the memory is inferred from both the retrieved internal information and external information to form a (memory) belief. Finally, the rememberer decides whether or not to communicate and how to communicate the belief to others. In this view, an autobiographical belief is the summative evaluation of the truth status of the remembered events at the time of retrieval (see also Otgaar et al., 2014; Scoboria et al., 2014, Scoboria & Henkel, 2020). Notably, since the information available could change in each instance of remembering, the truth status assigned to the events could therefore also be revisited and substantially altered, possibly resulting in the change of memory statements such as retraction of allegations of sexual abuse (Blank, 2017; Ost, 2017)." (Page 4 Line 58-85)

Q18. Second, on p6 the idea of a decision criterion is introduced, but a high degree of knowledge about SDT is assumed here. I don't think that SDT needs to be explained in full, but at least a brief explanation of what the criterion is doing, and a few more examples of why it might move in response to different circumstances would help the reader to follow the logic of the experiments. Given that the experiment is making some fairly fine distinctions between different types of rating that sound similar (but aren't) this detail is important so that readers can fully understand the difference between the ratings, and how they each relate to the criterion.

Response:

Thanks for this comment. We briefly introduced SDT at the beginning in the paragraph in the revised manuscript. It now reads as follows:

"A more nuanced picture of the relationship between memory distrust and memory appraisal was revealed through analyses using the Signal Detection Theory (SDT, Green & Swets, 1966). SDT is a framework to analyze people's ability to differentiate between signal and noise (i.e., sensitivity or discriminability) and their thresholds to decide whether a given stimulus is a signal or noise (i.e.,

response criterion) in decision-making processes such as recognition memory. For example, some people may have a very conservative response criterion and only judge a stimulus as signal when the evidence is strong, resulting in fewer correct recognition of signal (i.e., hits) and fewer false recognition of noise (i.e., false alarm). Others, however, might respond more liberally, judging a stimulus as signal given moderate evidence and having more hits and more false alarms." (Page 6 Line 117-125)

Minor Clarity points:

Q19. The words "filler" and "new" seem to be used interchangeably in the method, with no description of how they relate to each other. Provided that the terms "old" and "new" are well described, I see no reason why the word "filler" needs to be used at all, which will help with readers following the method.

Response:

Thanks. We have changed target/filler to old/new through the manuscript.

Q20. How long did participants have to complete the second part of the experiment? What's the potential range of the gap between the first and second part?

Response:

Thanks for this important question.

Based on our previous experience with online platform such as Connect and Prolific, it is reasonable to expect that the data collection for each session will be completed in a few hours. For example, in our previous experiment, it took less than 1 hour to recruit 120 participants on Connect.

In the current study, we expect to complete session 1 data collection within 3 hours. For the second session, we expect a long time (e.g., 6 hours) to complete data collection given the pool will be restricted to participants who have signed up for Session 1. The potential range of gap between Sessions 1 and 2 therefore will be 21 (24-3) hours to 30 (24+6) hours.

To limit the variability of the gap between encoding and later tests, we will close the Session 1 3 hours after the first signup regardless if the planned sample size has been met. If the number of participants is smaller than the planned sample size, we will then run another cycle of data collection until the sample size is met.

In the revised manuscript, we added the following information in the procedure section:

"In the current study, we expect to complete session 1 data collection within 3 hours. If, however, there are not enough participants signing up for the study within a 3-hour window, we will close the signup for Session 1 after 3 hours and run another cycle of data collection until the planned sample size is met." (Page 13 Line 270-273)

"Twenty-four hours later, Session 2 will be made available online. Participants will then have a 6-hour window to sign up for and complete Session 2, after which the session will be closed." (Page 13 line 276-277)

Further, we added the estimated completion time and range of gap in a footnote (Footnote 4).

Q21. Exclusion criteria – This strikes me as an experiment which requires participants to be both naïve as to the purpose of the experiment, and also at least to some degree uncertain as to whether their beliefs are accurate. Thus, in the exclusion criteria, it would be appropriate to consider ceiling and floor effects, as well as participant's responses to the open question about what they thought the experiment was about.

Response:

Thanks for this important comment. We added one more exclusion criterion based on the open-ended question regarding the purpose of the experiment.

"4) Participants who indicate, in the open question regarding the goal of the experiment, that they distrusted the feedback they received will be excluded from the analyses. Two independent coders (Y. Zhang and one research assistant) will code participants' responses regarding this exclusion criterion." (Page 17 Line 357-360)

In summary, I think this experiment should be conducted, but the bar lowered for how much existing knowledge is required to understand the design. I look forward to seeing the results.