Review report

*Managing Disclosure Outcomes in Intelligence Interviews* is a stage-2 registered report focusing on empirically testing the proposed model of disclosure decisions in intelligence interviews. In their preliminary study, the authors found that unguarded information with low cost and high benefit were most likely to be disclosed while guarded information (high cost and low benefit) were most unlikely to be disclosed. Rates of disclosure for low-stake (low cost and low benefit) and high-stake (high cost and high benefit) information fell in between. In the planned study, the authors reported failure to find support for the Cost\*Benefit interaction (an average effect in GLMM) found in the preliminary study. The authors speculate that this inconsistency resulted from the heterogeneity of participants’ risk and benefit perceptions.

I have only a few comments on the stage 2 report:

1. Both in the analyses of the preliminary study and the replication, the authors reported the ICCs of unconditional model (referring to the models without fixed effects I think) and consider the ICCs of items represents the individual differences in participants’ risk/ benefit perception of the item contents. In my opinion, this is only valid, if for every item, it can be paired with low stake/ high stake/ guarded/ unguarded based on probabilities. Based on my reading, this seems to not be the case. Instead, the items were fixed to be in certain conditions but not others. Perhaps it would be more appropriate to report the ICCs of items in the conditional models (which controls the experimental manipulation by including them as fixed effects) as the individual variation in risk/benefit perception of item contents.
2. The authors seem to attribute to the non-significant result of the replication study to the individual variations in the perception of the contents (resulting in weak manipulation/ larger noise). I find this a valid explanation as well. However, this does not seem to explain the fact that with a similar ‘weak manipulation’, there was a significant result in the preliminary study. Moreover, the ICC of items reported in the unconditional model was larger than in the replication study, which could suggest a great individual variation in the former case (or conversely, a stronger manipulation, given we couldn’t separate these two in an unconditional model, see my point 1).
3. If we look at the descriptive stats in the replication study, it is still consistent with the model’s prediction, which could be of value to mention in the discussion.
4. Also, it seems that compared to the preliminary study, the differences in disclosure rates between conditions was smaller in the replication study. Perhaps the authors could compare the two studies and discuss what could be the reasons behind this.
5. One way to avoid the large individual variation in risk/benefit perception is to measure participants’ risk and benefit perception of these items directly and use these as predictors in the model. Of course, then we are not directly manipulating risk/benefit (even if we manipulate the factors, the ratings are a product of manipulation by trait interaction), thus losing ‘power’ for causal inference. Nevertheless, it could be a complementary approach to the current design.

For example, in one of my previous paper examining memory verification strategy (Zhang et al., 2023, also see Nash et al., 2017), we asked participants to rate the perceived cost and benefit of each strategy and use these ratings to predict the likelihood of using such strategy. If the authors agree with my point, perhaps this could be added in the discussion section.

(the authors need not to cite either my paper or the paper of Nash et al., 2017)

**Reference**

Nash, R. A., Wade, K. A., Garry, M., & Adelman, J. S. (2017). A robust preference for cheap-and-easy strategies over reliable strategies when verifying personal memories. Memory, 25(7), 890-899.

Zhang, Y., Nash, R. A., & Otgaar, H. (2023). Preference for cheap-and-easy memory verification strategies is strongest among people with high memory distrust. Memory, 31(7), 978-988.

1. For the concluding remarks, perhaps it could be useful for readers if (in)consistent results from the current study are briefly mentioned as well, instead of only ‘what is needed to be done in the future’.

I hope that the authors find the my comments helpful and I apologize in advance if I misunderstood the MS in my reading.

Yikang Zhang