- overclaiming/inaccuracy: “replication” not done. When added, some results do not in fact repliate

- misinterpreation: non significant result interpreted as evidence for null. Although, additional analyses suggest that the magnitude of the effect is very small, suggesting that the effect may indeed be null within a SESOI.

- misinterpreation/overclaiming: claims re moderation, specifically an “equal and opposite” patterns of effect, are tested via both being significantly difference from zero and descriptively in opposite directions, no direct test of moderation. Visual inspection suggests they are “opposite” but not “equal”.

- threats to validity: M-ratio measure did not report reliability. When bootstrapped, it was .64 in study 1 and .19 in study 2. Given that the reliability of the measures involved places an upper limit on the correlations that can be observed between them, this raises the concern that the results here are too good to be true, perhaps due to overfitting of the factor scores which would inflate associations.

- threats to validity: On the one hand, factor structure was shown to be relatively stable after cross-validation, although results from study 1 are not robust. However, on the other hand, Fried’s recent general critique of the logic of higher-order factor models of psychopathology issues some stern warnings regarding the alluring yet invalid results that such analyses can provide, and how it can stymie scientific progress.

- auxhiliary hypotheses: While it goes beyond the question of error and into the domain of legitimate scientific debate, I would argue that Roualt et al.’s boldest claim is the one that it does not emphasise: that it has discovered that there is a latent factor structure of psychopathology. This serves as an auxiliary hypothesis for the actual hypotheses testsed, ie whether metacognitive evaluations are associated with these latent factors. The substantive claims regarding metacognitive evaluation depend entirely on the validity of the latent factors, but the research does not engage with the substantive literature on the latent factor structure of psychology, whether it is a sensible question (which is contentious), and if so what the specific factor structure is (which is equally contentious).

Dear Ian,  
  
Thanks to your input, it became clear to me that — although the study is technically very well reproducible — there are problems regarding the validity. Although I think there is no overfitting problem, there are inconsistencies between Experiments 1 and 2 that were not reported. In light of additional reanalyses that you suggested, some of the study's claims seem to be unsupported now.  
  
# Mismatch Between Experiment 1 and 2 Regarding CIT Factor  
  
I reinvestigated the match between Experiment 1 and 2 in multiple ways. All lead to the same results: Experiment 1 only partially replicates the results of Experiment 2 with regard to the relation between the factor analysis and metacognitive performances.  
  
- I performed a regression of the perceptual task performance measures on the questionnaire scores in Experiment 2 (these were only mentioned for Experiment 1 in the study). I replicated what the study reported: Anxiety and depression questionnaires are associated with low mean confidences, the obsessive-compulsive questionnaire is associated with high mean confidences. But while anxiety and depression also showed the negative association with mean confidence in Experiment 1, compulsiveness did not have a positive association in Experiment 2 (if there is any, it is negative).  
  
- I took the factor loadings of Experiment 2 and used them to compute scores for the participants from Experiment 1. Again, the results are mixed. The anxious-depression factor (AD) still has a negative regression coefficient. But the positive regression coefficients for the compulsiveness factor (CIT) on mean confidence disappear. Instead, an effect of the social withdrawal factor (SW) on mean confidences emerges.  
  
- Finally, I investigated overfitting in the factor solutions. I appreciate that you made me aware of the problems that can arise when plugging unvalidated factor scores into regression analyses. To check if their results were misleading and only due to factor solution ambiguity, I performed a cross-validation: On one part of the data set I constructed the factors and matched them to the factors of Experiment 2. On the other part of the data set, I computed regression coefficients. (I repeated this multiple times in a 10-fold cross-validation.) Results are again as above: Regression coefficients from Experiment 2 survive this more appropriate method. But Experiment 1 again shows a mixed pattern: The regression coefficients for AD replicate, but the negative effect of CIT disappears and a negative effect of SW appears.  
  
Taken together, I think there is no problem of overfitting in this data set. There is simply enough data. But there is a problem with inconsistent data patterns between Experiment 1 and 2 after doing the same analyses on both. The CIT factor effect of Experiment 2 simply vanishes in Experiment 1 (it almost points in the opposite direction).  
  
# Constructs Reliability and Validity  
  
I don't think there is a problem of reliability of the factor analysis results but there is a problem of reliability of the M-ratio measure.   
  
Regarding the factor analysis, I did it for Experiment 1 as you suggested. See an overview in the appended file, factor\_loadings\_exp\_1\_and\_2.pdf. The factor solution is surprisingly consistent between the experiments. Also, the authors validated their factors with those of Gillaan et al. (2016) in Figure S9. So I see no problem with this — assuming of course that the pool of questionnaires is reasonable. Unfortunately, I cannot judge that conclusively. But it seems reasonable to me.  
  
Following your question, I computed reliabilities of their perceptual task measures. I did not do split-half reliabilities because I don't know of an appropriate Spearman-Brown-correction for M-ratio, which may not behave as naively expected. Instead, I did Monte Carlo bootstrapping (Pronk et al., 2022), which may overestimate reliabilities (Kahveci et al., 2022). I found relatively high reliabilities for accuracies and mean confidences (Rel > .9) except for accuracy in Experiment 2, which is to be expected because of the staircase procedure reducing heterogeneity (Rel = .5). Reliabilities of M-ratio were low with .64 in Experiment 1 and even lower in Experiment 2 with Rel =  .19 due to the staircase procedure. Note again that the Monte Carlo approach tends to overestimate reliabilities. As another side note, I also changed my analyses to match theirs better. Previously I used a different estimation procedure for M-ratio. Because I now switched to their referenced Matlab code the problem does not come from my implementation.  
  
Given these results and in combination with the fact that I failed to reproduce reliable effects on M-ratio, I now have doubts regarding the claimed associations with this measure of metacognitive efficiency.  
  
# Overclaiming  
  
I agree now with your evaluation that some claims are overstated. Note that the statement: "Together, these results reveal that the AD and CIT symptom dimensions exert equal and opposite effects on two key aspects of metacognition" (p. 447) doesn't mean that the effect on confidence is equal and opposite to the effect on metacognitive efficiency (the latter is clearly smaller). The authors mean that the effect of AD on these measures is equal and opposite to the effect of CIT on them. But this is also problematic: They didn't test for equality as you said and, based on the reanalysis results above, their Experiment 1 doesn't support this. The effect of CIT is very unclear.  
  
I don't see a problem regarding their claims of a null effect with task performance: "[...] despite accuracy and parameters governing decision formation remaining unaffected" (p. 447). Although, this is indeed the dreadful practice of interpreting no underlying effect from a non-significant result, the sample size is simply large enough here. I added a reanalysis showing that one standard deviation change in any of the factor scores or questionnaire scores changes the accuracy by less than 1% (limits of 95% confidence intervals).  
  
# Conclusion  
  
I have revised my review (attached) accordingly. I do believe now that there is a problem with interpretations about the CIT factor due to inconsistencies between their two experiments. In consequence, interpretations are too strong as well.  
  
I am aware that I should have found these problems myself in the first round. I was focused too much on the technical replication and paid too little attention to the content claims. I appreciate that you brought these points up. I did not mention your contribution to this error-finding process in my updated review because I did not know whether that would have been appropriate. But I think the credit belongs to you.  
  
Let me know if you have other concerns or thoughts regarding this review.  
  
Best,  
Sascha  
  
  
Kahveci, S., Bathke, A., & Blechert, J. (2022). Reliability of reaction time tasks: how should it be computed?. PsyArXiv. [https://osf.io/preprints/psyarxiv/ta59r](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fosf.io%2Fpreprints%2Fpsyarxiv%2Fta59r&data=05%7C02%7Cian.hussey%40unibe.ch%7C3ad16bee5049483e192d08dcb8b12630%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638588318507803990%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=1A2q%2Fa1Ct4%2FOXMMdffwIHG%2FsPXNq4q4tl047SbEbW50%3D&reserved=0)  
  
Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. Psychonomic Bulletin & Review, 29(1), 44-54. [https://doi.org/10.3758/s13423-021-01948-3](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.3758%2Fs13423-021-01948-3&data=05%7C02%7Cian.hussey%40unibe.ch%7C3ad16bee5049483e192d08dcb8b12630%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638588318507814421%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=56d7ZkHT%2FTCJKsLuUuzRM67jFsqWE94rntJng8xoVfo%3D&reserved=0)