

Reply to Gillan et al.: What should be the null hypothesis in computational psychiatry?

Noam Sarna^{a,1}, Reuven Dar^a, and Matan Mazor^{b,c}

We thank Gillan et al. (1) for this exchange. To clarify, response biases and statistical artifacts are certainly not the only possible interpretation of existing data linking metacognitive biases and self-reported psychopathology. Substantive associations remain a viable alternative, and one that should be tested in future studies. Our main point is rather that, going forward, positive claims of such associations should rest on evidence that goes beyond what is expected from these nuisance factors alone.

In the case of response biases, we believe that current evidence does not meet this bar. The assertion of low inattention rates in the PIP (2) and Neureka (3) datasets rests on low inattention *detection* rates (PIP, 8.3%; Neureka, <1%), which are indeed well below the 24% we detect (4) and the 22% Zorowitz

et al. (5) find in a clinical sample. These lower rates are likely affected by the detection method: PIP used instructed-response items ("if you are paying attention, select x"), a poor measure

Author affiliations: ^aSchool of Psychological Sciences, Tel Aviv University, Tel Aviv 6997801, Israel; ^bAll Souls College, University of Oxford, OX1 4AL Oxford, United Kingdom; and ^cDepartment of Experimental Psychology, University of Oxford, OX1 3EL Oxford, United Kingdom

Author contributions: N.S. and M.M. designed research; N.S. and M.M. performed research; N.S. analyzed data; and N.S., R.D., and M.M. wrote the paper.

The authors declare no competing interest.

Copyright © 2026 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution License 4.0 (CC BY).

¹To whom correspondence may be addressed. Email: noamsarna@mail.tau.ac.il.

Published October 16, 2026.

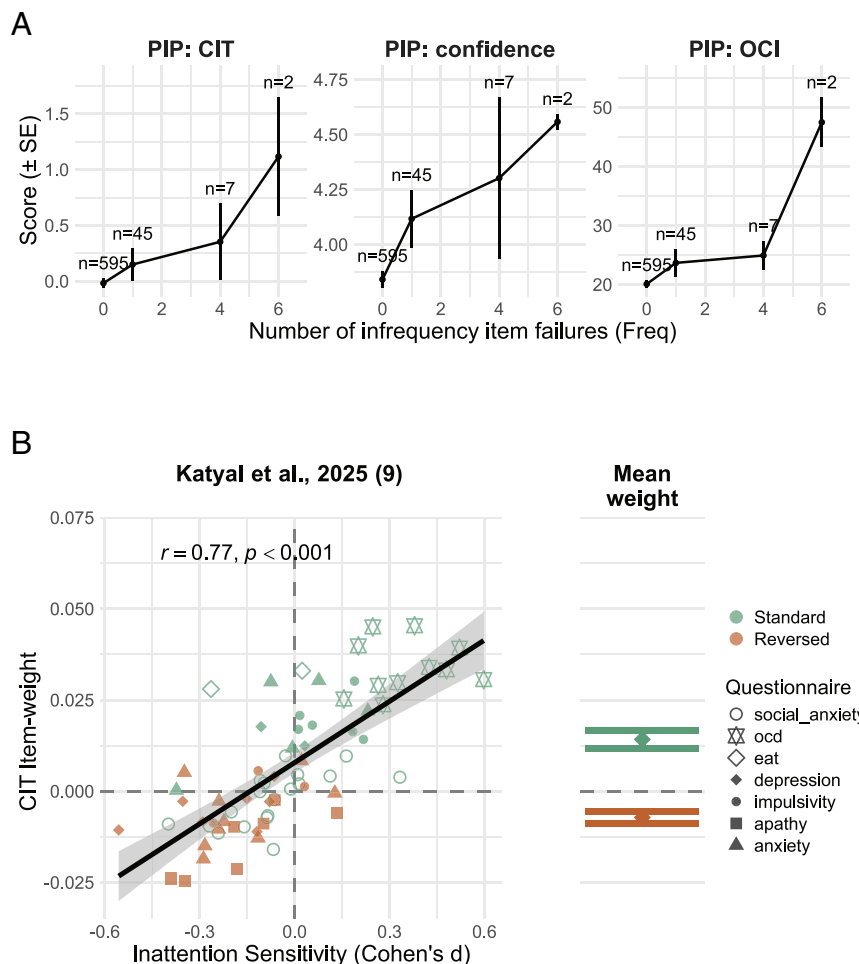


Fig. 1. (A) Inattentive participants in PIP (2) show elevated CIT, confidence, and OCI-R scores. Adapted from Sarna et al. (4, SI Appendix, figure A17). (B) The effect of reversed items and inattention on the CIT factor in Katyal et al. (9). Each point is a questionnaire item (shape = questionnaire; color = standard vs. reversed). x-axis: item sensitivity to inattention (Cohen's d , inattentive vs. attentive); y-axis: CIT item-weight. Item weights are assigned to reversed items after reversal. Diamonds (Right) show mean $\pm$ SEM weight for standard (green) and reversed (orange) items. Reversed items are used as safeguard against acquiescence. However, the CIT systematically assigns reversed items negative weights, effectively turning this safeguard into an additional source of bias. Adapted from Sarna et al. (4, figure 4), which is licensed under CC BY-NC-ND 4.0.

of inattention (6, 7), and Neureka a single infrequency item (in our study (4)), a well-fitting model showed that a single infrequency item detects only 39% of inattentive participants, leaving 61% undetected; SI Appendix, figure A15). Because these confounds are far larger than the effect of interest [roughly by a factor of 10; (4, reply to Gillan, SI Appendix, section 2)], and since participants flagged as inattentive in the PIP study display exactly the profile attributed to compulsive individuals (elevated CIT, OCI-R, and confidence; Fig. 1A), the CIT–confidence association in these datasets remains consistent with the effect being entirely driven by residual response biases (8). This is not the only possible interpretation of the data, but it remains a viable one.

The same holds for covariate control. The emergence of a significant confidence–OCD association when controlling for a correlated covariate (depression) is compatible with a spurious correlation produced by a collider bias. Beyond a simulation, we showed in our own data that this mechanism produces a correlation between measures that are independent by construction: acquiescence and SDS, a questionnaire with 50% reversed items (4, reply to Gillan, SI Appendix, section 1). Similar to the case of response

biases, whenever the observed pattern is compatible with a likely alternative explanation, we contend that this should serve as the null hypothesis to reject.

Finally, regarding the differences between CIT and OCI-R, we note that correlations with confidence can be driven by acquiescence, independent of inattention. While the OCI-R is affected by acquiescence only due to the absence of reverse-coded items, CIT is doubly susceptible: it includes reverse-coded items but systematically undoes the semantic reversal, assigning the already reversed scores a negative weight (Fig. 1B). This again makes differences in correlations with confidence difficult to attribute to psychopathology alone.

Gillan et al. (1) point to important next steps to address these open questions, and we agree. Specifically, new designs and measures that are less vulnerable to these confounds will more unequivocally tease apart substantive associations from the effects of response biases and statistical artifacts.

Data, Materials, and Software Availability. Analysis code and data have been deposited in pnas-response (<https://github.com/self-model/pnas-response/>) (10). Previously published data were used for this work (2, 9).

1. C. M. Gillan, S. M. Fleming, M. Rouault, T. X. F. Seow, Disentangling biases in responding from substantive effects in computational psychiatry. *Proc. Natl. Acad. Sci. U.S.A.* **123**, e2624329123 (2026).
2. C. A. Fox et al., An observational treatment study of metacognition in anxious-depression. *eLife* **12**, RP87193 (2023).
3. C. A. Fox et al., Reliable, rapid, and remote measurement of metacognitive bias. *Sci. Rep.* **14**, 14941 (2024).
4. N. Sarna, R. Dar, M. Mazar, Biased and inattentive responding contribute to apparent metacognitive biases in mental health. *Proc. Natl. Acad. Sci. U.S.A.* **123**, e2520832123 (2026).
5. S. Zorowitz, J. Solis, Y. Niv, D. Bennett, Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nat. Hum. Behav.* **7**, 1667–1681 (2023).
6. A. J. Barends, R. E. De Vries, Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personal. Individ. Differ.* **143**, 84–89 (2019).
7. D. J. Hauser, N. Schwarz, Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods* **48**, 400–407 (2016).
8. J. Westfall, T. Yarkoni, Statistically controlling for confounding constructs is harder than you think. *PLoS One* **11**, e0152719 (2016).
9. S. Katyal, Q. J. Huys, R. J. Dolan, S. M. Fleming, Distorted learning from local metacognition supports transdiagnostic underconfidence. *Nat. Commun.* **16**, 1854 (2025).
10. N. Sarna, R. Dar, M. Mazar, Code and data for “Reply to Gillan et al.: What should be the null hypothesis in computational psychiatry?” (2026). <https://github.com/self-model/pnas-response/>. Accessed 27 September 2026.