Review of PCI Registered Reports #149 "Taking A Closer Look At The Bayesian Truth Serum"

Summary

The RR aims to replicate and extend a recent study by one of the authors (Schoenegger, 2021). That paper demonstrated an impact of Bayesian truth serum (BTS) incentives on (online) respondents' answers to experimental philosophy questions, but left open whether the impact was due to the incentives themselves or to certain other features of the BTS protocol. The RR proposes four between subject treatments: One is (almost) a direct replication of the 2021 paper; the remaining three treatments test different possible explanations for the impact of BTS incentives.

Evaluation

The RR addresses valid scientific questions with clearly stated hypotheses. I have no concerns about methodology, sampling and proposed analysis. However, I would recommend another iteration in the design, addressing the following issues.

- The RR protocol replaces the original 2021 instruction text "Recent work by researchers at MIT that has been published in the academic journal Science..." with the less specific: "Recent work by researchers that has been published in leading peer-reviewed journals..." I don't have strong feelings about which version is better, this is an empirical question. However, the change makes the RR not strictly speaking a replication. A "failure to replicate" per Row 1 in the Design Template could be attributed to a change in instructions.

- The authors are missing an opportunity to test whether the original result replicates *without* predictions. That is, Information Scores (which support truth-telling incentives) can be computed even for respondents that do not make predictions. In principle, therefore, one could elicit predictions from only a few 'holdout' respondents, or, alternatively, for only some questions presented to each respondent. The practicality of the BTS method would be enhanced if the burden of making predictions was reduced or eliminated (for most respondents). The current Prediction condition tests whether predictions are sufficient; adding a condition without predictions but with the BTS instructions cover story would resolve whether predictions are even necessary.

- I understand why the authors wish to retain the original seven items used in (Schoenegger, 2021). However, the pattern of results in the original study suggests that BTS incentives do not affect the distribution of answers for philosophical problems with moral / responsibility / virtue content, but do affect problems with the (arguably more challenging) knowledge / truth / causality content. One interpretation is that in the former case, respondents have robust prior intuitions that drive their answers whether or not they are under incentives. With the less familiar problems in the latter set, careful reading of the question may be more critical, leading to a different level of comprehension and distribution of answers under incentives. If so, then the problems with moral content are not the best domain to test for incentive impact.