Report on

"Taking A Closer Look At The Bayesian Truth Serum: A Registered Report"

This Registered Report aims to test the performance of the Bayesian Truth Serum (BTS): an incentive scheme for truthful elicitation of subjective answers. It builds on previous work by the authors, which has shown that questions incentivized by the BTS show a different distribution of answers than those elicited without. The aim of the current work is to disentangle whether these differences are due to the theoretical properties of the BTS, the presence of incentives, or the predictive power.

Overall, I found the proposed study interesting and well motivated (although see below). The document is transparent and mostly easy to read. Nevertheless, I have some comments on the outline of the document as well as the proposed analysis that warrant some revisions:

1. The first comment is procedural. I found the material necessary to evaluate the proposal to be somewhat scattered. The tests are mentioned in the introduction, its sequence is discussed the "potential results" section, and the proposed statistical tests are found the methods section. Piecing this together was a bit of work. I would propose a structure that is more traditional (at least in my field of behavioral economics). First, use the design/methods section to describe the experiment. Then, in a a separate hypothesis section, specify explicit and numbered hypotheses, framed in terms of observable data patterns. Finally, for each hypothesis specify exactly what data it will be applied to (e.g. only those vignettes where there is a significant difference on an previous hypothesis test), which test you will conduct, and as a part of this, what evidence will count as a confirmation of the hypothesis. While this is mainly rearranging of materials already present in the report, it should clarify the exposition.

2. My second point is more substantial. The report proposes to first establish differences between the BTS and the no-incentive condition. Then, for any vignettes that show a statistically significant difference, the analysis will look at differences between BTS and the remaining conditions (Prediction and Additional Money), to conclude whether the overall difference can be explained by the subcomponents of the BTS.

   I see two problems with this procedure. First, the null-hypothesis of the proposed non-parametric tests is that the distribution of answers is the same. The rejection of the null hypothesis therefore does not say anything about the nature of the difference or the direction of the change. Thus, it is theoretically possible that you find a difference in your

first comparison (BTS vs. No Incentives) that goes in one direction, and a between BTS vs. Additional Incentives that goes in the opposite direction, both times with statistical significance. In this case, the conclusion that the first difference is driven by the second, is the exact opposite of what one should conclude. Maybe this is an extreme / unlikely scenario, but many variations are possible, e.g. the first test might be positive because of increased variance in the data and the second because of a shift in central tendency.

There is a related problem in the sequencing of the analysis. For the same reason as highlighted above, it is possible that the first comparison might not statistically significant (BTS vs. No Incentives), while there are statistically significant differences between "No Incentives" and "Prediction" or "Additional Incentives". This would suggest that the combined features of the BTS reverse or ameliorate some effects of the individual features. Again, this may not be very likely, but it cannot be ruled out ex-ante, and it would not be picked up by the analysis. The analysis also rules out the identification of an overall (across vignettes) effect of incentives or prediction integration, which seems a pity from a scientific perspective.

The core problem here is that the implicit assumptions about the nature of the effect that remain untested by the very general null hypothesis of the proposed non-parametric tests. To overcome this problem it may be wise to consider additional analysis, like the use of hierarchical regression models. This might also be a way to get at an overall effect of different treatments, by combining the different vignettes. For the latter, one should of course use appropriate multilevel techniques like random effects to account for dependence between observations from the same vignette or experimental subject.

3. Finally a smaller point: The motivation misses a large literature in economics on the role of incentives in experiments and surveys, see e.g. Schlag et al. (2015). Even if this literature focuses mostly on the elicitation of objective events, it is relevant for some of the claims made in the opening paragraphs. I also note that one of the authors is in the same institute as a prominent BTS theorist, whose work goes uncited (Baillon 2017, Baillon et al. 2020.)

References

Baillon, Aurelien, Han Bleichrodt, and Georg Granic. "Incentives in surveys." mimeo, 2020.

Baillon, Aurelien. "Bayesian markets to elicit private information." Proceedings of the National Academy of Sciences 114.30 (2017): 7958-7962.

Schlag, Karl H., James Tremewan, and Joël J. Van der Weele. "A penny for your thoughts: A

survey of methods for eliciting beliefs." Experimental Economics 18.3 (2015): 457-490.