14 March 2025

Dear Romain,

Re: Does Truth Pay? Investigating the Bayesian Truth Serum with an Interim Payment

Thank you for your thoughtful and encouraging feedback on our Stage 2 Registered Report. We greatly appreciate the careful consideration you and the reviewers have given our work. In response to your comments, we have revised the manuscript, as detailed below.

1. **Deviation from Stage 1 Protocol**

   Recommender's Comment:

   "If the registered test is feasible, I would agree with Philipp and would recommend you to put it in the manuscript as the main test, presented as the registered procedure. If there was an internal contradiction and the test was not feasible, as I understood, I would agree with your deviation. Here, to address Philipp's concern, you could possibly run the registered test (with Welch adjustment) on each of the five datasets and report the associated p-values (either in Supplementary Materials or in a footnote). This would ensure that the two closest alternatives to the original (infeasible) test are presented."

   Philipp's Comment:

   "While the authors originally stated that they would apply a Welch adjustment to account for variance inequalities (page 17 in Stage 1 manuscript), the final analyses use HC3 robust standard errors instead (described on page 19 in Stage 2 Manuscript and applied on page 23). Although HC3 can be a suitable alternative, the authors should adhere to their preregistered plan by reporting the Welch-adjusted results as the primary analysis, exactly as specified. The HC3-based analysis can then be presented as an exploratory robustness check.

This approach would fully honor the original preregistration and simultaneously demonstrate the robustness of the findings through additional analyses."

Response:

As noted, our preregistered approach specified a Welch adjustment to address variance inequalities. However, the method required pooling variance estimates across five imputed datasets, which, we confirm, was not feasible within the planned contrasts framework. Accordingly, we applied HC3 robust standard errors as a suitable alternative (Long & Ervin, 2012) and reported this deviation in the manuscript.

To address Philipp's concern, we have now conducted Welch-adjusted t-tests separately on each imputed dataset, as you suggested, to ensure that the two closest alternatives to the original (infeasible) test are presented. These results are reported in the Supplementary Materials (Table S0). The manuscript has been updated on pages 18 and 22 accordingly. Importantly, the Welch t-tests produced results consistent with the primary analysis, confirming the findings reported in the manuscript.

2. **Interpretation of Bayes Factors**

Sarahanne's Comments:

"I think you still need to explain why you reject fixed thresholds entirely"

"You still refer to Hoijtink et al. (2019b) as a guide for interpreting your Bayesian analysis, and still fail to properly specify which aspects of that extensive work you are following… You don't clarify (for example) whether you are incorporating posterior probabilities as an alternative or supplement to Bayes factors. A sentence specifying exactly how you apply Hoijtink's framework would strengthen this section"

Response:

We have clarified in the manuscript (pp. 19-20) that the specific guidance we are citing from Hoijtink et al. (2019b) is from the section entitled *How Large Should the Bayes Factor Be?* (p. 545). This section argues against using fixed threshold values for Bayes factors, cautioning that such thresholds can introduce the same issues seen in NHST, such as publication bias and arbitrary decision-making. Accordingly, we have updated the manuscript (p. 20) to state:

"... we considered Bayes factors as direct and quantitative indicators of the evidence for (or against) the alternative hypothesis in comparison to the null hypothesis rather than applying strict thresholds"

Additionally, while posterior probabilities provide an intuitive interpretation of evidence strength, we have opted not to report them in this case, given the supplementary status of the Bayesian analysis and the lack of a strong basis for specifying prior probabilities. To make this clearer, we have added the following statement to the results section (pp. 23-34):

"Given the supplementary nature of this analysis and the absence of a strong basis for prior probabilities, we did not convert the Bayes factors to posterior probabilities but interested readers could do so by multiplying the Bayes factors by their own choice of prior odds."

3. **One-Tailed Tests and Backfire Effects**

Recommender's Comments:

"You used here a one-tailed test, which, as Sarahanne underlines, is not appropriate to discuss potential backfire effects. I believe that you made the best use of the Registered Report format by using one-sided tests. Indeed, the theory was very clear about why we should have expected an improvement in answers with the BTS, so it was clear that the tests needed to be one-sided to maximize statistical power. You find some evidence for backfire effects: we can see it as new exploratory evidence that can be used in a confirmatory analysis by a future

Registered Report. The confirmatory investigation did not aim to discuss backfire effects in the first place."

Sarahanne's Comments:

"Given that the first Bayes factor is extraordinarily large, you could have reflected on whether this indicates strong evidence for a backfire effect rather than just a failure of BTS. A discussion of how the Bayesian results fit into your broader conclusions is what I suggest here"

"Another issue concerns your decision to preregister a one-tailed test – the main result would have been statistically significant had you preregistered a two-tailed test rather than a one-tailed test, but you don't acknowledge this as a potential limitation. Since you used a one-tailed test, you implicitly assumed a directional effect in advance – that is not an ideal choice when testing an intervention that could plausibly backfire, I would argue. Some reflection on whether this decision was appropriate and/or how it may have influenced your conclusions would add some value to the discussion."

Response:

Our decision to preregister one-tailed tests was guided by a strong theoretical rationale for expecting less socially desirable responses under the BTS mechanism and to maximise statistical power. However, to reinforce the intention of our approach, we have updated the manuscript (p. 22) to state: *"As a one-tailed test was preregistered, this result is interpreted within that framework"* while acknowledging that the result would have been significant with a two-tailed test.

While the one-tailed tests were appropriate for the confirmatory analysis in determining whether our directional hypotheses were supported, we acknowledge that they were not suited for detecting backfire effects. In this regard, the Bayes factors provided more useful insights. To better reflect this, we have expanded the discussion on page 31 to clarify that these results may indicate a backfire effect rather than merely a failure of the BTS:

"While the primary confirmatory analysis did not aim to test backfire effects, the supplementary Bayesian analysis identified an unexpected pattern that could indicate increased social desirability bias under the BTS".

Additionally, we have clarified that the Bayesian findings should be interpreted as part of an exploratory analysis and that future research should formally investigate backfire effects using a preregistered confirmatory approach. To this end, we have added the following statement as part of the limitations section of the discussion (p. 31):

"These exploratory findings suggest that future preregistered studies should not only address how well social desirability assumptions align with participant norms but also investigate potential backfire effects using a targeted confirmatory approach".

4. **Alternative Analytical Approaches**

Sarahanne's Comments:

"... you don't consider whether a different analytical approach (e.g., mixture models or priors that allow for unexpected effects) might provide additional insights. Instead, you speculate about the reasons for the observed pattern without critically evaluating whether your chosen statistical framework was well-suited to detecting and interpreting such an effect. I think this point needs some attention in the write-up."

Response:

As noted above, our analytical framework was designed to test our preregistered directional hypotheses and was well-suited for evaluating the expected effects of BTS. Accordingly, it was not specifically designed to detect unanticipated effects, such as potential backfire effects. Given the exploratory nature of these findings, we agree that future confirmatory research could benefit from alternative analytical approaches, such as mixture models or priors that account for unexpected effects. To reflect this, we have updated the manuscript

(p. 31) to suggest that future studies should consider employing suitable analytical approaches to further investigate the potential backfire effect.

We trust that these revisions appropriately address the feedback. We look forward to your response and hope our submission is now suitable for recommendation.

Best regards,

Claire