Peer Community In Registered Reports

Do interim payments promote honesty in self-report? A test of the Bayesian Truth Serum

A recommendation by ***Romain Espinosa*** *based on peer reviews by* ***Sarahanne Miranda Field*** *and* ***Philipp Schoenegger*** *of the STAGE 2 REPORT:*

Claire M. Neville, Matt N. Williams (2025) Does Truth Pay? Investigating the Effectiveness of the Bayesian Truth Serum with an Interim Payment: A Registered Report. OSF, ver. 3, peer-reviewed and recommended by Peer Community in Registered Reports. https://doi.org/10.31219/osf.io/s3znc

Submitted: 15 January 2025, Recommended: 08 April 2025

Cite this recommendation as:
Espinosa, R. (2025) Do interim payments promote honesty in self-report? A test of the Bayesian Truth Serum. *Peer Community in Registered Reports*, 100979. 10.24072/pci.rr.100979

Published: 08 April 2025

Surveys that measure self-report are a workhorse in psychology and the social sciences, providing a vital window into beliefs, attitudes, and emotions, both at the level of groups and individuals. The validity of self-report data, however, is an enduring methodological concern, with self-reports vulnerable to a range of response biases, including (among others) the risk of social desirability bias in which, rather than responding honestly, participants answer questions in a way that they believe will be viewed favorably by others. One proposed solution to socially desirable responding is the so-called Bayesian Truth Serum (BTS), which aims to incentivize truthfulness by taking into account the relationship between an individual's response and their belief about the dominant (or most likely) response given by other people and then assigning a high truthfulness score to answers that are surprisingly common (Prelec, 2004).

Although valid in theory (under a variety of assumptions), questions remain regarding the empirical utility of the BTS. One area of concern is participants' uncertainty regarding incentives for truth-telling – if participants don't understand the extent to which telling the truth is in their own interests (or they don't believe that it matters) then the validity of the BTS is undermined.

In the current study, Neville and Williams (2025) tested the role of clarifying incentives, particularly for addressing social desirability bias when answering sensitive questions. The authors administered an experimental survey design (N=877) including sensitive questions, curated from validated scales, that are relevant to current social attitudes and sensitivities (e.g., "Men are not particularly discriminated against", "Younger people are usually more productive than older people at their jobs"). Three groups of participants completed the survey under different incentive conditions: the BTS delivered alone in a standard format, the BTS with an interim bonus payment that is awarded to participants (based on their BTS score) halfway through the survey to increase certainty in incentives, and a Regular Incentive control group in which participants receive payment without additional incentives.

The authors analyzed the effectiveness of the BTS through two registered hypotheses. First, the authors found that the BTS did not increase agreement with socially undesirable statements (compared to the control group), as theory would suggest, and even observed an opposite effect. This result, which could be confirmed by follow-up studies, raises some concerns about the robustness of the BTS method. Second, the authors conjectured that introducing an interim payment in the BTS mechanism would help reinforce its credibility in the eyes of the participants and would thus magnify its effect. However, the authors failed to detect a statistically significant difference between the standard BTS and interim-payment BTS mechanisms. Overall, the results of Neville and Williams (2025) call for some caution in the use of the BTS and for further work to better understand the contexts in which the BTS might be a useful tool to mitigate social desirability in surveys.

This Stage 2 manuscript was evaluated over one round of in-depth review by two expert reviewers and a second round of review by the recommender. After the revisions, the recommender judged that the manuscript met the Stage 2 criteria and awarded a positive recommendation. **URL to the preregistered Stage 1 protocol:** https://osf.io/vuh8b **Level of bias control achieved:** Level 6. *No part of the data or evidence that was used to answer the research question was generated until after IPA.* **List of eligible PCI-RR-friendly journals:**

- Advances in Cognitive Psychology
- Advances in Methods and Practices in Psychological Science
- Collabra: Psychology
- Experimental Psychology
- In&Vertebrates
- Meta-Psychology
- Peer Community Journal
- PeerJ
- Royal Society Open Science
- Social Psychological Bulletin
- Studia Psychologica
- Swiss Psychology Open

References:

1. Neville, C. M & Williams, M. N. (2025). Does Truth Pay? Investigating the Effectiveness of the Bayesian Truth Serum with an Interim Payment: A Registered Report [Stage 2]. Acceptance of Version 4 by Peer Community in Registered Reports. https://doi.org/10.31219/osf.io/s3znc

2. Prelec, D. (2004). A Bayesian truth serum for subjective data. Science, 306, 462-466. https://doi.org/10.1126/science.1102081

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.31219/osf.io/s3znc
Version of the preprint: 3

Authors' reply, 14 March 2025

Download author's reply
Download tracked changes file

Decision by Romain Espinosa, posted 25 February 2025, validated 25 February 2025

Minor revision requested

Dear authors,

Thank you very much for submitting your Stage 2 to PCI-RR. I was very pleased to see your completed study and look forward to recommending it.

Two of the three referees who reviewed your Stage 1 gratefully accepted to also examine your completed study. As you will see, both referees have a very high opinion of your work.

Considering their feedback and my own reading of your work, I would like to invite you to (slightly) revise your manuscript before issuing the final recommendation.

Here are some comments:

- First, I agree with both referees that you did an excellent job in conducting your analysis as you committed to in the Stage 1. Congratulations!

- Second, I tend to concur with Philipp's opinion regarding the deviation from the Stage 1 protocol. Deviations happen, and let me say that they are completely normal because we cannot anticipate everything that is going to happen. You made your deviation very transparent and I am grateful for this. Nevertheless, I think that we should try to stick as much as possible to the original plan to maximize the credibility of this submission format (that's the philosophy at PCI-RR at least). If I understood correctly, the registered method could not be implemented because of the number of datasets you are working on. It follows that there was a contradiction in the registered document, which makes the test infeasible. Is that correct?

If the registered test if 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.

- Third, I understand from Sarahanne's review that the interpretation of the Bayes Factor is a delicate task, especially when it comes to defining decision thresholds. From my perspective, I think that you made it clear that this analysis would not determine whether you reject the null hypothesis, which makes it more like a tool to discuss the previous findings than a test that is part of the confirmatory analysis. I agree with Sarahanne that you amended your Stage-1 to drop the (arbitrary) thresholds. That being said, I believe that Sarahanne's comment might reflect general concerns in your discipline and that you might be willing to take the opportunity to clarify this issue with one or two additional sentences (e.g., specifying what you use from Hoijtink et al., as Sarahanne suggests). The decision is yours.

- Fourth, 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. That being said, here also, I believe that Sarahanne's view might be shared by other colleagues in your discipline, and that you might want to add a few words to make these points clear(er) to your audience. Here again, the decision is yours. I am looking forward to receiving your revised manuscript, which should not take too long.

Thanks again for choosing PCI-RR for your work. I really see your work as a very neat application of Registered Reports, with a high contribution to the discipline (in terms of method for the BTS) and a rigorous implementation of the RR procedure.

Thanks again for trusting us with your work at PCI-RR,

Best regards,

Romain

Reviewed by Philipp Schoenegger, 16 January 2025

The Stage 2 manuscript follows almost all of the preregistered steps and reports the outcomes in a well-justified manner, with no unwarranted changes from the Stage 1 manuscript.

However, I believe there is one notable deviation from the preregistered protocol that is not properly addressed. Specifically, 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.

Reviewed by Sarahanne Miranda Field, 24 February 2025

Dear authors,

Thank you for your Stage 2 Registered Report. I appreciate the thorough execution of your study and your careful interpretation of the findings, but I am not completely happy with how you have engaged with my suggestios in the last round of reviews if I am honest. I therefore have a few more comments regarding your handling of the Bayesian analysis and its integration into the results and discussion sections. These primarily relate to the interpretation of Bayes factors, the specificity of your reference to Hoijtink et al. (2019b), and the implications of your statistical choices. Just a note about comparing between stage 1 ad 2, I think you remained methodologically consistent with the Stage 1 plan, but that the interpretation of results and discussion of limitations could have been more thorough and that's what I am focusing on the most here as it also comes back to some of the original comments I made.

First, you removed the problematic statement about "values greater than 1 supporting the corresponding hypothesis," which I think is good, though now you simply state that Bayes factors will be interpreted contextually following Hoijtink without actually explaining why you avoid thresholds. I find this problematic. While you're avoiding the mistake of implying that BF10 > 1 always provides meaningful evidence, I think you still need to explain why you reject fixed thresholds entirely rather than (e.g.) highlighting BF10 > 3 as moderate evidence which is (largely) accepted practice. Obviously, we can deviate from what's typically done, but whatever choices are made must be motivated sufficiently, especially when it comes to

Second, 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. While you now mention using a Cauchy prior and pooling across imputations, you dont 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, and will better satisfy the concern I originally had for the stage 1 RR.

In the results, you performed a Bayesian analysis but did not fully integrate it into the interpretation (in my opinion). You state that a BF10 of approximately 24,757 supports a non-null effect, while BF10 ≈ 0.103 supports the null, but do not really discuss the practical implications of these findings which I think is quite an oversight. 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 – I think it would be an addition that would really make your findings more informative.

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.

Finally, while your discussion on the possible backfire effect is interesting, 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.

In general, I judge this to be a strong and well-executed stage 2, and I appreciate the effort that went into conducting and reporting it. My comments are intended to help clarify your Bayesian analysis and strengthen the discussion of your findings - I hope my tone doesn't come across as negative in any way. I look forward to seeing your revisions!

I sign all of my reviews,
Dr. Sarahanne M. Field