30 August 2024

Dear Dr. Espinosa,

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

Thank you for the comprehensive and thoughtful feedback from each reviewer on our Stage 1 manuscript and for your overview as the recommender, which has guided us in consolidating the various comments. We were delighted to find general support for our study and the research questions we pose.

In the paragraphs below, we first respond to the overview you provided before addressing each piece of feedback. You will see that we have taken on board the reviewers' principal concerns regarding multiple hypothesis testing and potential null effects and have decided to adjust our analysis strategy to address these concerns and reflect other aspects of the feedback received. We hope that these adjustments satisfy the issues raised.

## Recommender's Overview

> **First, I share Philipp' and Sarahanne's concerns about the interpretation of null results. The absence of evidence is not the evidence of absence. Here, equivalence testing or smallest-effect size of interest testing are your friends if you want to go further in this direction. Second, there is a concern about the multiple hypothesis testing that Philipp, Martin, and I felt uncomfortable with. I read Rubin (2021) that you cite and I understand your comment: you test each hypothesis separately with separate tests, so you could argue they are not part of the same family of hypotheses. However, your overall conclusion ('overarching inferential criteria' section) combines the results of these hypotheses. So, here, it is clear that they are part of a family of hypotheses (wrt to your conclusion). A complex question is how to address the interdependence between the tests. Third, this then raises**

**questions about the sample size, as Martin underlines, once you have addressed the previous issue(s).**

**As far as I am concerned, I see no problem with sticking to a frequentist approach. Bayesian approaches can be informative in case of null results, but there are also tools in frequentist statistics. I think that the exclusion of some observations poses no problem but it might be important to think about potential issues (which I might not have anticipated). In my discipline, imputation is rarely done, so I would not push you in this direction if you do not find it appealing (except, here again, if I miss important elements).**

We have taken into account the reviewers' concerns about multi-hypothesis testing and null effects and their suggestions to incorporate Bayesian methods. Accordingly, we have updated our analysis strategy by retaining a frequentist method that addresses multiple hypothesis testing concerns for the primary analysis supplemented by Bayes factors.

For the primary analysis, we will use planned contrasts, as suggested by Martin Schnuerk. This method is well-suited to our research questions, enabling specific, theory-driven comparisons between groups. While orthogonal contrasts help ensure that each hypothesis test is independent, there remains a possibility that the familywise Type I error rate could be inflated beyond the nominal 5%. To mitigate this risk and address the reviewers' concerns about multiple-hypothesis testing, we will implement a Bonferroni correction, adjusting the alpha level for each of the two planned contrasts to $\alpha = 0.025$.

Regarding the interpretation of null results, while equivalence testing or SESOI testing could be effective tools in this context, we have opted to incorporate Bayes factors as a supplementary analysis. This approach aligns with the contemporary interest in Bayesian methods (Wagenmakers et al., 2017) and ensures that our findings remain informative even when the primary analysis

does not yield significant effects. It is important to clarify that we are not specifying null hypotheses in our primary analysis; rather, Bayes factors will provide additional insights into potential null effects without influencing the primary inferential criteria.

Finally, regarding missing data, we have updated our strategy to employ a multiple imputations approach in accordance with Rubin's (1987) guidelines. While missing data is likely to be minimal and, consequently, deletion might not introduce bias, we acknowledge that multiple imputation is methodologically more robust. Accordingly, we have made this adjustment to our strategy to ensure that our data analysis is as rigorous and reliable as possible.

We hope our revised analysis strategy adequately addresses the main concerns raised. The following paragraphs respond to each of the individual points made by the reviewers, including our response to other important identified matters, such as the calculation of i-scores. We also provide brief clarifications on the more minor points contained in the review.

## Romain Espinosa's Comments

- **1.1 Page 6: the sentence about Menapace and Raffaelli (2020) is unclear.**

The reference to Menapace and Raffaelli's (2020) findings was intended to illustrate how the BTS has been shown to reduce, though not entirely eliminate, hypothetical bias (primarily due to socially desirable responding) in contingent valuation studies. However, based on the feedback received, we acknowledge that the reference to overlapping coefficients was unclear. Given that the reference was not critical to our narrative and considering our word limit for target publications, we have decided to remove it altogether.

- **1.2 How do you take into account the fact that receiving an intermediary payment might convey information about whether one underestimates or overestimates other participants' views or strategies?**

The score upon which the payment of the interim bonus will be determined is an aggregate of each participant's score for each of the five items that make up Part 1 of the survey. These items cover different constructs and will be presented in random order. As described in more (mathematical) detail in response to 3.2a) below, the BTS score calculated for each participant's response to each item consists of both an information score (i-score) and a prediction accuracy score. The i-score measures whether the participant's answer is surprisingly common (or not) by comparing the actual frequency of the participant's answer with the collective prediction of the distribution of that answer. Therefore, it is only the prediction accuracy component of the combined score that measures an individual participant's accuracy in predicting the frequency of a particular answer.

The formula for the prediction accuracy score measuring how well a respondent $r$'s prediction of the distribution for answer $k$ matches the actual distribution of responses is:

$$\alpha \sum_{k} \bar{x}_k \, log \left( \frac{y_{kr}}{\bar{x}_k} \right)$$

Where:

- $\alpha$ is a constant that fine-tunes the weight given to the prediction error.
- $\bar{x}_k$ is the actual average frequency of answer $k$ given by all respondents.
- $y_{kr}$ is respondent $r$'s prediction of the distribution for answer $k$.

Since only the prediction accuracy component measures prediction accuracy and the interim bonus payment is based on the aggregate score, we think it will be challenging for a participant to discern any useful information about whether they have accurately predicted other participants' scores. Moreover, participants will not have visibility of the calculation method unless they click on the link we provide in the information sheet to Prelec's (2004) original paper, which we anticipate is unlikely for most participants. Even if they do read the paper, we still think it would be difficult for most participants to use the feedback provided by receiving an interim payment to ascertain their predictive accuracy on a combination of individual items to improve their scores in the second part of the survey.

- **1.3 Hypothesis table: I do not understand why H2 is part of the second research question. It is clear that H1 is part of RQ1, and H3 is part of RQ3. However H2 explores BTS-IP compared to no BTS, so it seems to be closer to RQ1.**

  As covered in the response to the recommender's overview above, we have adjusted our analysis strategy to include planned contrasts. Accordingly, the study hypotheses have been restated as follows:

  - H1: Participants subjected to the BTS (with or without an interim payment) will have significantly higher mean scores indicating agreement with socially undesirable statements compared with those in the Regular Incentive condition.
  - H2: Participants subjected to the BTS with an interim payment will have significantly higher mean scores indicating agreement with socially undesirable statements compared with those subjected to the BTS alone.

  There is now a very clear connection between the research questions (reproduced below) and our hypotheses, which is apparent in our updated design planner table (page 10 of our revised manuscript).

- RQ1: Can the BTS effectively incentivise honesty in Likert scale questions prevalent in psychology research?
- RQ2: Does the inclusion of an interim payment enhance the efficacy of the BTS mechanism?

- **1.4 It seems to me that the hypotheses are not independent. For instance, if H1 is rejected (BTS increases scores), but H2 is not rejected (BTS-IP has a positive but non-significant impact), we cannot reject H3.**

  As noted above, the hypotheses have been restated and will be tested using a planned contrasts approach, addressing the issue of the hypotheses' interdependence.

- **1.5 If we want to compare BTS and BTS-IP, we should in BTS say that their payment will be £0.25 for Part 1 and £0.25 for Part 2, but they will know that at the end, no? (If people are risk averse, they should prefer two smaller lotteries than one large lottery to reduce risk, no?)**

  We had not initially viewed the situation this way, as our focus was on the dependent variable (DV)—the social undesirability score—and whether the BTS mechanism has a significant effect and not on which condition participants prefer. However, we acknowledge that it is rational for participants to prefer two smaller lotteries. Accordingly, we have updated our approach (as indicated on pages 12 and 13 of the revised manuscript). In the BTS condition, participants will be informed that BTS scores will be calculated separately for each part, with payment provided after completing both parts. In accordance with our revised bonus amounts (see our response to 4.2 below for more details), participants could earn:

  - £0.00 if they do not rank in the top 50% for truthfulness in either part;
  - £0.50 if they rank in the top 50% in either Part 1 or Part 2 but not both;

- £1.00 if they rank in the top 50% in both parts.

This approach effectively creates two smaller lotteries (one for each part), which participants may prefer as it allows them to potentially secure part of the bonus earlier (after Part 1). However, the actual payment will be made only after the completion of both parts. This method aligns with the preference for smaller, independent lotteries while maintaining the integrity of the BTS mechanism.

- **1.6 Why run 2-tailed tests if the predictions are directional?**

While our hypotheses are directional, we initially opted for two-sided t-tests to capture any potential backfire effects. This approach was informed by Field (2018), who recommends using two-sided t-tests to maintain a conservative stance and avoid overlooking unexpected results. As noted by Aron et al. (2023), there is ongoing debate among researchers about whether one-tailed tests should be used even when there is a clearly directional hypothesis. Nevertheless, after considering all the reviewers' feedback, we believe that specifying one-sided tests using planned contrasts is a more suitable primary analysis approach for the present study, complemented by Bayes factors to gain additional insights into potential null effects. This adjusted approach is reflected in our updated manuscript (pages 17 and 18).

**1.7 Please put the fourth bullet point in « inferential criteria » in an « outcome neutral test » paragraph.**

This change has been actioned (page 19).

- **1.8 I would also suggest adding some outcome-neutral tests for ceiling and floor effects given that you use a 5-point Likert scale. (Statistical power might be very small in case you have a large mass of the distribution on one of the extreme values of the Likert-scale in the control group.)**

In our study, the DV is the summed Likert scale responses across all main survey items, with higher scores representing more socially undesirable answers. Given that the DV is an aggregate score, we believe floor and ceiling effects are less likely. Aggregating responses across multiple items should mitigate the risk of extreme distribution skewing, maintaining sufficient variability for meaningful analysis. Analysing effects within specific subsets is beyond this study's scope, which we have clarified in the manuscript (see footnote on page 15). Accordingly, we do not plan to perform outcome-neutral tests for these effects.

- **1.9 Regarding the conclusions to be drawn (« overaching inferential criteria »). For the 1st bullet point: Note that failing to reject H3 does not induce that the BTS-IP is not more efficient. As the motto says: « the absence of evidence is not the evidence of absence ». If you want to conclude on this, you should define a smallest-effect size of interest (SESOI) that you pre-register and, ex-post, see if you can reject the fact that the difference between BTS-IP and BTS is smaller than this difference. (We discuss one possibility under the name of Sequential Unilateral Hypothesis Testing, section 3, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4110803).**

  As indicated earlier, we have taken on board the collective feedback regarding the evaluation of null effects and adjusted our analysis strategy accordingly. Both frequentist and Bayesian methods offer suitable tools for this purpose. Although we found the Sequential Unilateral Hypothesis Testing approach interesting, we have opted for Bayes factors, which allow us to examine null effects while aligning with the contemporary interest in Bayesian methods among researchers. It is important to clarify that our primary analysis is not designed to test null hypotheses; rather, it focuses on detecting directional effects. In this context, the use of Bayes factors serves as a supplementary analysis to provide additional insights into potential null effects, helping to clarify the strength of evidence against the null hypothesis without influencing the primary inferential criteria. This

approach aligns with our decision to avoid rigid interpretation thresholds for Bayes factors, as we aim to use them contextually to enhance our understanding of the data rather than to draw definitive conclusions about null effects.

- **1.10 I think that the final decision could be illustrated with a decision tree. (See attached for an example from what I understood.)**

  Thank you for providing the decision tree illustration. However, given that our primary analysis has been streamlined to focus on planned contrasts testing only two hypotheses, the inferential criteria have become more straightforward, as outlined on page 17 of the revised manuscript, making a decision tree less necessary in this context.

## Phillip Schoenegger's Comments

- **2.1 First, with respect to participants, is there a specific reason why the authors aim to collect responses from six nations instead of one? Prolific will easily be able to provide you with the required sample from just one country without introducing this potential (albeit unlikely) confound. Additionally, within Prolific, are you going to use further selection criteria? E.g., with respect to the number of successfully completed tasks, language competency, etc.**

  We acknowledge that Prolific can provide a sufficient sample from one country. However, our decision to collect responses from six nations aims to highlight the international scope of the research, with the countries selected having similar language and cultural backgrounds. Given our use of random assignment to conditions, the selection of multiple countries cannot plausibly introduce confounds.

  Regarding selection criteria within Prolific, we have updated our approach based on your feedback and Dr. Schnuerk's feedback on our exclusion rate. We will

include fluent English proficiency and a minimum of 20 prior completed Prolific surveys as pre-screeners. The manuscript has been revised to reflect these additional requirements (page 9).

- **2.2 Second, I would ask the authors to clarify the hypotheses and the tails of their analysis. I think stating the null hypotheses (instead of the directional hypotheses) should improve clarity and also set up the authors to say more about what kind of conclusions they would draw in the case of a backfire effect (i.e., if the BTS makes responses less 'honest'). The paper can still include directional predictions, of course, but as it currently stands there seem to be a couple of potential patterns of data that are not accounted for.**

  Thank you for suggesting that we consider using null hypotheses instead of directional hypotheses. While we recognise the value of this approach, especially for two-tailed tests, we have decided to retain directional hypotheses in our primary analysis. This choice allows us to express our predictions clearly and simply without unnecessarily expanding the number of hypotheses.

  To address concerns about potential null and backfire effects, we have supplemented our analysis with Bayes factors to test non-directional alternatives. This approach allows us to assess significant differences between groups regardless of direction, providing a more comprehensive interpretation of the data. We have also incorporated an exploratory analysis of response distributions, as described in our response to item 2.4, to further examine potential backfire effects and other unaccounted patterns in the data.

- **2.3 Third, I strongly suggest you adjust your p-values in a way that's distinct from your current approach, where the adjusted level is still at 5%. I leave it up to the authors to decide an adjustment procedure that they think is appropriate, but I strongly suggest you pick a method that does not**

**result in de facto non-adjusted p-values, as I am not convinced by Rubin (2021) and its application to your design.**

As noted above, we have revised our analysis strategy to use a planned contrasts approach for the primary analysis, as suggested by Martin Schnuerk. This method suits our research well, allowing for specific, theory-driven comparisons between groups based on prior expectations. Planned contrasts (Ψ) will compare mean scores (μ) between the following groups:

- Ψ1: BTS (with or without interim payment) vs. Regular Incentive
- Ψ2: BTS with interim payment vs. BTS alone

Weights will be assigned as follows:

- Ψ1: -2 (μRI) + 1 (μBTS) + 1 (μBTS+IP)
- Ψ2: 0 (μRI) - 1 (μBTS) + 1 (μBTS+IP)

While the orthogonality of these contrasts ensures that each contrast tests a distinct hypothesis (confirmed by the sum of the products of the weights equaling zero, i.e., (-2*0) + (1*-1) + (1*1)), there remains a possibility that the familywise Type I error rate could be inflated beyond the nominal 5%. To mitigate this risk and ensure the robustness of our findings, we will apply a Bonferroni correction (Bonferroni, 1936), adjusting the alpha level for each of the two planned contrasts to $\alpha = 0.025$.

**2.4 Fourth, while I respect the authors' choice in focusing on mean differences as their main outcome variable, I would suggest using at least one test of differences in response distributions as an exploratory additional analysis (that need not play into the main hypotheses but should be reported). I leave it up to the authors to pick whatever test they think is most appropriate, but I think some type of analysis may help provide a better understanding of a potential null effect.**

In response to the suggestion to explore differences in response distributions, we will conduct Chi-Square tests of independence for individual items, with responses cross-tabulated by condition, as an exploratory analysis. We agree that these tests could provide insights into potential null effects identified in the planned contrasts. This exploratory analysis will be included in the supplementary materials. It will not influence our primary analysis, which focuses on the social undesirability score (the sum of item responses) as our DV rather than individual item responses. The Analysis Strategy has been updated accordingly (page 19 in the revised manuscript).

- **2.5 Fifth, as far as I can tell, the authors do not provide a plan on how they deal with null effects. The preregistered outcome interpretations talk of a failure of hypothesis support (which is framed as a directional effect), so a backfire and a null effect would be treated identically, which doesn't seem right to me. This gets back to the framing of hypothesis as directional having them stated as 'no difference' null hypotheses allow you to also consider scenarios where effects do not work (established, for example, by some type of equivalence tests with the bounds drawn from the effect size used for the power calculation) or actually backfire.**

  We appreciate your concern regarding the treatment of null and backfire effects. In response, we have refined our analysis strategy to include planned contrasts for primary hypothesis testing, supplemented with Bayes factors to evaluate non-significant findings. This combined approach allows us to differentiate between null and backfire effects more effectively. Furthermore, as noted above, we have incorporated an exploratory analysis to look at differences in response distributions, which will be included in the supplementary materials.

- **2.6 Sixth, you say that "the BTS will be tested for its ability to elicit truthful responses across various dimensions". I might have missed this, but won't you be aggregating across all question types? If so, how will you**

**test for the effects in subsets of these questions? Will these be additional analyses? If so, these p-values also have to be adjusted and their relationship to the main hypotheses stated.**

Thank you for pointing out that this statement might be misleading. You are correct that the responses to the survey items will be tested in aggregate. The intention was to highlight that the items encompass a range of constructs, not to imply that responses for each construct will be analysed as subsets. Analysing effects within specific subsets is beyond the scope of the current study. We have clarified this in the manuscript (see footnote on page 15) to ensure our primary focus on aggregate data is clear.

- **2.7 Seventh, why are you preregistering a Welch's t-test without knowing the variances in both groups? Would it not be better to preregister the criterion by which you will decide on the test (which may well end up being a Welch's t-test)?**

  Initially, we pre-registered Welch's t-tests based on guidance from Delacre et al. (2017), who recommend using Welch's t-tests by default to address potential unequal variances between groups. Following this peer review, particularly Martin Schnuerk's suggestion, we have updated our primary analysis approach to planned contrasts with a Welch adjustment. This approach maintains the same rationale as our initial plan (Delacre et al., 2017) and is further supported by Zimmerman (2010), who advocates for avoiding a two-step procedure, such as conducting Levene's test (Levene, 1960) first, which could inflate Type I error rates. Page 17 of the manuscript has been updated accordingly.

- **2.8 Eighth, how are the questions presented to participants? Are they randomised?**

In all conditions, each main question is paired with a prediction question. The question pairs for items 1-5 in Part 1 and items 6-10 in Part 2 will be randomised within their respective parts. Our intention to present questions in random order is now clearly communicated on page 12.

- **2.9 Ninth, how will you include demographic variables in your analyses (if at all)?**

  The inclusion of demographic variables in our statistical analyses is beyond the scope of the current study, which we have now stated more clearly in the manuscript (see 'Descriptive Analysis' section, page 17). Given our use of random assignment to conditions, statistical controls for demographic variables are not necessary to avoid confounding. However, all the data we collect, including demographic data (age, gender, education level), will be shared on the Open Science Framework (OSF), allowing other researchers to extend the analyses should they wish to explore potential demographic effects.

- **2.10 Tenth, I would like the authors to also reflect on their choice of items, as some seem to be from work published in the 90s (and as such conducted prior to that). Is it not very plausible that some of the uncomfortable and/or sensitive topics of the 90s are quite distinct from the ones today? For example, discourses of gender and race/ethnicity have changed dramatically, and so has what is seen as sensitive, potentially putting into question your assumptions for the outcome variables.**

  Your comments about the evolution of social attitudes and sensitivities since the 1990s, particularly in the areas of gender, race and ethnicity are well taken. While our primary focus was on selecting questions that had the potential to elicit SDR without causing personal upset, we agree that using contemporary scales ensures that the topics addressed in our questionnaire are relevant to current social contexts and reflect modern sensitivities. Accordingly, we have updated

our questionnaire to include items from scales developed or revised in the 21st century. Specifically, we have selected questions from the Social Dominance Orientation (SDO7; Ho et al., 2015) to replace the previous questions selected from the SDO6 (Pratto et al., 1994) and have updated the remainder of our survey with questions from the Belief in Sexism Shift (BISS; Zehnter et al., 2021) and the Succession, Identity and Consumption Scale of Prescriptive Ageism (SIC; North & Fiske, 2013). Examples include: "An ideal society requires some groups to be on top and others to be on the bottom" (SDO7), "Nowadays, men don't have the same chances in the job market as women" (BISS), and "Doctors spend too much time treating sickly older people" (SIC). We trust this update addresses your concern and enhances the validity of our outcome variables by aligning them with today's discourses on social dominance, gender-based inequalities and ageism. The manuscript has been updated to capture these changes on page 15.

- **2.11 Eleventh, some small things. I don't think abbreviations in the abstract are necessary (page 1). Early on (page 3), you state that the BTS is 'novel', though I am not sure this is entirely true given the amount of work that has been done on it already. What do you mean by 'overlapping coefficient' (page 6)?**

  We note your view regarding the use of abbreviations in the abstract. Our decision was influenced by the stringent 200-word limit imposed by one of our target journals. Using the full terms 'Bayesian Truth Serum' and 'socially desirable responding' multiple times would significantly reduce the space available to convey the study's key features. In the revised abstract of the Stage 2 manuscript, we will aim to limit abbreviations to 'Bayesian Truth Serum' if possible.

  Regarding the term 'novel' used to describe the BTS, we recognise that it has been the subject of considerable research since its introduction by Prelec (2004). We intended to highlight its innovative approach relative to traditional methods in

eliciting truthful responses. However, we agree that this term may be misleading in representing the BTS as new and have removed it.

The reference involving the term 'overlapping coefficient' has now been replaced—please see our response to 1.6 above.

## Martin Schnuerk's Comments

- **3.1 Justification of hypotheses: I wonder, however, whether interim payments could also have a negative effect on some participants. Assume that a participant provides, at least in their view, truthful responses. Yet, based on the i-score ranking, this person is not among the top group and receives no bonus payment. They may conclude that the algorithm to detect truthful responding isn't working properly or that the feedback is corrupt, potentially resulting in decreased trust in the procedure. The authors clearly don't expect this effect and they may have good reason for this – however, I recommend that they spell out these reasons more explicitly and justify their expectations/hypotheses more carefully.**

  We appreciate the reviewer's thoughtful feedback regarding the potential negative impact of interim payments on some participants, particularly the concern that participants who believe they are providing truthful responses but do not receive an interim bonus payment may lose trust in the BTS mechanism.

  We consider it unlikely that interim payments will reduce trust in the BTS mechanism. This expectation is based on findings from Weaver and Prelec (2013), which tested a version of the BTS mechanism where participants were exposed to dynamically calculated i-scores as they progressed through the survey. Findings showed that participants adjusted their behaviour to be more truthful in response to feedback on their i-scores and related earnings. We have clarified this rationale in the revised manuscript (page 8) to ensure our expectations are clearly justified.

However, we acknowledge the possibility that some participants might perceive the BTS mechanism as faulty if they provide what they believe to be truthful responses but do not receive a reward. This could indeed result in less honest or effortful responding in Part 2 of the interim payment condition, potentially affecting the support for H2. Yet, the possibility that our hypotheses might not be supported is precisely why this research is necessary and interesting, as it allows us to test these hypotheses and explore whether interim payments influence participant behaviour as predicted.

- **3.2 a) Measures: The authors mention the i-score as their measure for ranking and rewarding participants. In the BTS, the i-score is usually combined with a prediction accuracy score. Why is this score not considered here? The reason for choosing one and not the other score is rather implicit at the moment. Also, from the verbal definition on p. 4, the exact calculation is not clear to me: Is this how the score is calculated for each item? Or for each response on each item? Or each respondent? Without proper mathematical notation that includes indices denoting respondents, items, and response options, this is unclear. Therefore, I recommend that the authors include a more explicit justification for their choice of measure and replace the verbal equation on p. 4 with a formally correct mathematical equation.**

  First of all, in response to this feedback, we have updated the 'Bayesian Truth Serum' section of the manuscript to clarify that the BTS functions at the level of individual questions (Prelec, 2004). Each response will be assigned a specific score (‘BTS score’) by combining an information score (i-score) and a prediction accuracy score. These BTS scores for each question will then be aggregated to provide a total score for each respondent.

Additionally, as suggested, we have included mathematical formulae to describe the calculation of the BTS score, as outlined in Prelec's (2004) paper. The mathematical notation specifies that the overall BTS score for respondent *r* for answer *k* combines the i-score and the prediction accuracy score as follows:

$$BTS\ Score \ = \ \sum_k x_{kr}\ log\left(\frac{\bar{x}_k}{\bar{y}_k}\right) \ + \ \alpha\sum_k \bar{x}_k\ log\left(\frac{y_{kr}}{\bar{x}_k}\right)$$

Where:

- $x_{kr}$ is 1 if respondent $r$ chooses answer $k$, and 0 otherwise.
- $\bar{x}_k$ is the actual average frequency of answer $k$ given by all respondents.
- $\bar{y}_k$ is the geometric mean of the predicted frequencies for answer $k$ made by all respondents.
- $\alpha$ is a constant that fine-tunes the weight given to the prediction error.
- $y_{kr}$ is respondent $r$'s prediction of the distribution for answer $k$.

- **3.2 b) Relatedly, are participants ranked within each condition or across conditions?**

  Participants will be ranked within each BTS condition to ensure independence between conditions. Those in the top 50% of scores in each condition will receive a bonus, avoiding any disadvantage to participants in the BTS condition compared to those in the BTS + IP condition. Ranking across conditions could introduce biases, especially since the interim payment mechanism in the BTS + IP condition is expected to influence scores. Treating each condition separately ensures a fair and accurate comparison, allowing for a clear analysis of the effects within each condition. This approach has now been clarified in the

manuscript (pages 12 to 14) as part of explaining the study procedures for each condition.

- **3.3 a) Statistical Analysis: First of all, I do not understand why the authors would conduct two-sided tests when the hypotheses are clearly directed. Directed hypotheses warrant one-sided tests. Otherwise, the statistical models at tests do not correspond to the substantive hypotheses, resulting in unnecessarily conservative tests. Moreover, and more importantly, the authors argue that the three tests are based on independent null hypotheses. I do not agree with this assessment because the hypotheses are clearly not mutually independent: H1 and H3 imply H2. Thus, the tests are, in fact, redundant and not independent (and neither are the corresponding null hypotheses). The authors formulate precise expectations about the ordering of mean scores across conditions (as noted above, see also #1). I suggest that these expectations be put to the test in a more critical and powerful way, namely, by means of planned contrasts.**

  While our hypotheses are directional, we initially chose two-sided t-tests to capture any potential backfire effects. This approach was informed by Field (2018), who advocates for two-sided t-tests to maintain a conservative stance and avoid overlooking unexpected results. As Aron et al. (2023) note, there is an ongoing debate about the appropriateness of one-tailed tests, even for clearly directional hypotheses.

  However, after considering all the reviewers' feedback, we believe that specifying one-sided tests using planned contrasts, complemented by Bayes factors, is a more suitable analysis approach for our study. Planned contrasts enable us to test our directional hypotheses more precisely and efficiently, aligning better with our specific expectations about the ordering of mean scores across conditions. This method increases the power of our tests without inflating our sample size

(Field, 2018). Furthermore, supplementing our primary analysis using Bayes factors addresses concerns about understanding potential null or backfire effects.

We thank the reviewers for guiding us towards this overall analysis strategy and believe that the revised approach effectively addresses the primary concerns raised.

- **3.3 b) Sample Size**

  As noted above, we have revised our primary analysis approach to use the suggested planned contrasts method. Consequently, we have updated our power analysis using G* Power to reflect these adjustments (see page 11).

- **3.3 c) Multiverse analysis: Another suggestion for the analysis plan is to complement the planned frequentist analysis with a Bayesian evaluation by means of Bayes factors.**

  Further to our previous responses, our revised approach includes a primary analysis using planned contrasts, supplemented by Bayes factors. This combination may be seen as aligning with the spirit of a multiverse approach. Although the supplementary Bayes factors analysis will not determine the support for the main hypotheses, it will provide additional context and help to better understand any potential null effects observed in the primary analysis.

- **3.4 Exclusions/Dropout: The authors adjusted their target sample size to account for potential exclusions. This makes a lot of sense, especially in online studies. However, I wonder whether the estimate of 5% may be too optimistic. It is based on Schoenegger's (2021) findings which, to my knowledge, were not based on 2 study parts and, thus, did not include additional dropout. From personal experience, dropout rates in multi-part studies can be quite high. Thus, the authors may want to increase the**

**proportion of participants to add to the target sample size to account not only for exclusions but also dropout between study parts.**

In light of this feedback, we revisited attrition rates in comparable multi-part studies to determine an appropriate allowance for both exclusion and potential dropouts. Kothe and Ling (2016) reported a 23% attrition rate between the first and second time points spaced one year apart in their online longitudinal study hosted via Prolific. In another Prolific-hosted longitudinal study with intervals spaced one month apart, Williams et al. (2024) reported a 13% attrition rate between the first and second time points. These examples suggest that longer intervals between parts can lead to higher attrition rates. Accordingly, we believe an exclusion rate of 10% for our study, with only a maximum break of 48 hours between parts, is reasonable. This rate, double that reported by Schoenegger (2021) for a single-part study, should adequately cover potential exclusions and dropouts. Page 11 of our revised manuscript has been updated to reflect this change.

- **3.5 Quality check—I found the wording of the quality check potentially misleading. It says, "What percentage of those with the highest information scores will receive a bonus?" 50% is the correct response. However, the phrase "those with the highest information scores" could be perceived as referring to a subset of participants, namely, those in the top 50%. In this case, the correct response would be 100% because all members of this subset should receive the bonus.**

  Thank you for highlighting the potential confusion in the original phrasing of the quality check. It has now been rephrased for clarity, with the relevant section of the manuscript (page 15) updated to read as follows:

  *"'At the end of Part 2, participants in each of the BTS conditions will be asked, 'What percentage of participants, ordered by their BTS scores, will be eligible for a bonus?' with options of 30%, 50%, or 100%."*

Please note that in the revised quality check question, the term "BTS score" replaces the term "information score" in accordance with our response to 3.2 a) above.

## Sarahanne Miranda Field's Comments

**4.1 I'm not convinced by your explanation for choosing a deletion procedure rather than multiple imputation procedure to handle missing data. While I think it's quite likely that missing data will be minimal, and that deletion is unlikely to introduce bias, why not just choose a method that is more methodologically robust? Can you explain your choice a little more, or change the strategy?**

The listwise deletion method was initially chosen due to the expectation of minimal missing data and, as a consequence, a low risk of introducing bias. However, we acknowledge that multiple imputation is methodologically more robust. In light of your feedback, we have updated our strategy to include using multiple imputations for handling missing data (see page 16 under 'Analysis Strategy'). This adjustment ensures that our data analysis is as rigorous and reliable as possible.

**4.2 I'm wondering about the payment amounts - they seem so small that I am wondering if they will be enough for the purpose. They may well be, of course, but I am not convinced because you (unless I missed something) do not provide a justification for why you have chosen these amounts and why you think they will be sufficient for their purpose. Please provide a clear motivation for the amounts chosen.**

Our study involves base payments for participation along with bonus payments under BTS conditions.

The base payments were informed by Prolific's guidelines, ensuring the amount converts to an hourly rate of £15 for survey completion, which compares favourably to the UK's National Living Wage of £11.44 (Gov.Uk, 2024). The manuscript has been updated on page 12 to provide this rationale.

The bonus amounts in our study were influenced by both successful applications of the BTS and budget constraints, balancing adequate power to detect a small effect size with financial limitations. Existing literature on the BTS shows a wide range of bonus amounts. For instance:

- Frank et al. (2017) offered US$0.50 to the top third of participants.
- Schoenegger (2021) provided £1 to the top third.
- Weaver and Prelec (2013) offered up to US$25.

Due to resource constraints, our initial sample size of 1,245 required a bonus payment at the lower end, with a maximum of £0.50 across both parts. However, by reducing our sample size to 876 participants using planned contrasts, we can feasibly increase the bonus to £1, as in Schoenegger’s study. This adjustment ensures participants could potentially double their overall pay, aligning incentives with other proven experimental applications.

While larger bonuses might yield better results, as suggested by Weaver and Prelec, they are beyond our budget. Nonetheless, previous studies indicate that even smaller bonuses can effectively elicit more truthful responses. Therefore, proceeding with our adjusted bonus payment is a practical and evidence-supported approach. This revised bonus strategy allows participants to potentially double their overall pay, which can be a significant incentive for online participants who often rely on surveys as a source of income (Peer et al., 2021).

The ‘Procedure’ section of the manuscript has been updated to reflect this revised bonus strategy; however, we stress that the feasibility of this approach is contingent on a sample size of around 876 participants.

**4.3 I did not find a clear justification for 10 questions. You say that you chose this number to keep it manageable, but why do you think 10 will be manageable, and not, say 15 or 20? Have you run a pilot to actually test the time cost to participants? 10 seems to be a low number, and without sufficient motivation about why you have chosen this number, I am left wondering if that will be enough questions.**

The justification for the number of questions is closely related to our rationale for the base payment. While there are ten main items, there are, in fact, 20 questions in total per participant when the predictions are included. Answering the prediction questions is quite involved as participants are required to estimate the percentage of others in that condition who chose each of the 5 Likert scale responses while confirming their predictions sum to 100.

When trialling the survey on volunteers outside the research team, the time required equated to a base payment of £1, which was at the limit of what we could afford based on the size of sample needed. Additionally, the time taken for participants to answer 20 questions aligns with the Prolific guidelines for fair compensation and ensures the survey remains manageable and engaging without causing fatigue (Denison, 2023).

However, the manuscript has been updated to clarify this decision and emphasise that the total number of questions and the involved nature of the prediction task justifies the chosen number to balance thorough data collection with budget constraints (see page 15 and 16). We believe this approach is both practical and justified, ensuring the integrity and feasibility of our study.

**4.4 What will you do if you get p-values greater than your cutoff alpha? Do you just shrug and go "well, we don't really know..."? Because that's about the extent of the information you will get if you use frequentist statistics. I *strongly* recommend using a Bayesian approach along side of frequentist (or even in place of!), because it allows us to deal with pro-null evidence and because Bayes factors can provide more information value than p-values can.**

We appreciate your suggestion, which has informed our updated analysis strategy involving planned contrasts supplemented by Bayes Factors.

We hope you'll agree that we have carefully considered the feedback you and the other reviewers provided. This feedback, we believe, has led to significant improvements in our study design and analysis strategy. We appreciate your guidance and look forward to your further comments.

Yours sincerely,

Claire Neville

# References


Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th ed.). Pearson.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commerciali Di Firenze*, *8*, 3–62.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92. https://doi.org/10.5334/irsp.82

Denison, G. (2023). *How much should you pay research participants?* Prolific. https://www.prolific.com/resources/how-much-should-you-pay-research-participants

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PLOS ONE*, *12*(5), e0177385. https://doi.org/10.1371/journal.pone.0177385

Gov.Uk. (2024). *National minimum wage and national living wage rates*. Gov.uk. https://www.gov.uk/national-minimum-wage-rates

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new

$SDO_7$ scale.. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Kothe , E., & Ling, M. (2019). Retention of participants recruited to a multi-year longitudinal study via Prolific. *PsyArXiv (OSF Preprints)*. https://doi.org/10.31234/osf.io/5yv2u

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford University Press.

Menapace, L., & Raffaelli, R. (2020). Unraveling hypothetical bias in discrete choice experiments. *Journal of Economic Behavior & Organization*, *176*, 416–430. https://doi.org/10.1016/j.jebo.2020.04.020

North, M. S., & Fiske, S. T. (2013). Act your (old) age. *Personality and Social Psychology Bulletin*, *39*(6), 720–734. https://doi.org/10.1177/0146167213480043

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*. https://doi.org/10.3758/s13428-021-01694-3

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037//0022-3514.67.4.741

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462–466. https://doi.org/10.1126/science.1102081

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. In *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Schoenegger, P. (2021). Experimental philosophy and the incentivisation challenge: A proposed application of the Bayesian truth serum. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00571-4

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, *50*(3), 289–302. https://doi.org/10.1509/jmr.09.0039

Williams, M. N., Ling, M., Kerr, J. R., Hill, S. R., Marques, M. D., Mawson, H., & Clarke, E. J. R. (2024). People do change their beliefs about conspiracy theories—but not often. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-51653-z

Zehnter, M. K., Manzi, F., Shrout, P. E., & Heilman, M. E. (2021). Belief in sexism shift: Defining a new form of contemporary sexism and introducing the belief in sexism shift scale (BSS scale). *PLOS ONE*, *16*(3), e0248374. https://doi.org/10.1371/journal.pone.0248374

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173–181. https://doi.org/10.1348/000711004849222