Dear Editor/Reviewers,

Thank you for the opportunity to revise this manuscript based on the comments outlined in detail below. In this letter, we provide all comments, our response to these comments, as well as an example of how we have implemented this change in our manuscript (if applicable), so that the time to go back and forth between this document and the revised manuscript is minimised.

Both reviewers indicated that the questions posed are scientifically valid and that the methodology is fitting. However, they proposed changes to the analysis plan and to the way the outcomes of these analyses are interpreted. In this letter, we address these comments, and where we have decided against implementing the suggestions, we have outlined our reasons.

Overall, every section of the previous manuscript has been reworked to some extent, while parts of the manuscript have been fully rewritten (see Revised Manuscript -Tracked Changes). The main changes that we have implemented are: (i) a more precise justification of expected effect size and sample size and (ii) a fully reworked and elaborated justification Potential Results/Implications section.

Smaller changes have been made throughout, focusing on readability and structure of the paper, correct contextualisation within the background literature, further justification of our design choices (if changed and if retained), all the while aiming to not have the manuscript balloon in size.

We hope that our revisions are satisfactory and that our manuscript can now be considered for in-principle-acceptance. However, should there been any further changes and amendment requested, we would gladly take them on board and revise our manuscript further.

Thank you very much again for your time and effort in helping us improve this manuscript!

EDITOR/ADDITIONAL REVIEWER:

Comment 1: Regarding the inferences based on results showing a non-significant difference between groups:
"The problem with the analysis plan is that it takes non-significance in itself as evidence against there being a difference. **What is needed is an inferential procedure that justifies a claim of no effect so that results could actually count against various predictions**. See https://psyarxiv.com/yc7s5/ for the typical alternatives and how to approach them: **power; equivalence tests of various sorts; Bayes factors.**
To justify a conclusion of no effect being there, there needs to be a scientifically motivated indication of what size effect there could be, if there were one. For power and equivalence tests this should be a minimally relevant effect. Thus, if the plan is to continue to use chi square then power is calculated with respect to Kramer's V or w. **A problem for the author to address is justifying the minimally relevant effect size. Then power can be calculated; and thus, according to Neyman Pearson, a non-significant result taken as grounds for asserting no difference**. No special power is required for PCI RR, but the power of each test should be known, if power is the tool used; but the author may wish to bear in mind that different PCI RR friendly variables may have requirements (albeit not RSOS). In any case, whatever the journal requirements, "no effect" cannot be concluded without justification.
One possibility, which the author may reject, is to say the BTS is useful in so far as it shifts the mean towards the truthful answer; thus a test of mean differences could be used. Power, equivalence testing and Bayes factors are all easier conceptually. One asks what shift of mean is of minimal interest (for power or equivalence tests); or what shift in mean, or range of shifts is scientifically plausible (for Bayes factors), which may be the easier question to answer."

Response 1: Thank you for pressing us on this! We now clearly state the expected effect size at Cramer's V=.1 based on the average previous effect being at V=.117, the smallest significant previous effect being V=.101, and a conventional 'small effect' being V=.1. We also added a disclaimer that there may be smaller effects in some contexts, which means that we will suspend judgement in cases of non-significance.

Example (p. 8):

We selected the expected effect size as follows: First, we averaged across all effects (both significant and non-significant) from the previous paper on the same items (Schoenegger, 2021). This yielded a mean Cramer's V=.117 as our expected effect size. (Note also that the smallest significant effect from the previous study was V=.101.) Further, as the standard 'small effect size' for Cramer's V is conventionally put at V=.1 and to be conservative, we will choose V=.1 as our expected effect size. As such, we will understand null effects as null effects up to this expected effect size and make this clear throughout the paper. Further, smaller effects may

be interesting in different contexts, which means that we will suspend judgement about whether there is a relevant effect for some contexts or not in cases of non-significant results.

Comment 2: 2. Regarding the planned comparisons between different groups and selection of items for these analyses:
"This [*refers to the planned comparisons*] is problematic because of a selection effect: By selecting extreme scores in the first comparison [*concerns the comparison between the BTS group and the main control group*], they will naturally tend to get differences in the second [*concerns the comparisons between the BTS and additional control groups*]. So **I would not do the pre-selection. The authors should just look at the overall evidence, for each comparison without preselection**."

Response 2: Thank you very much for weighing in on this. To be honest, we had this discussion just before submission and were torn on this very choice. We now follow your recommendation and do not preselect comparisons and have adjusted the manuscript as such.

Comment 3: Here, I had an additional question: given that with the two new control groups, the idea is to test to what extent alternative explanations can explain the base effect (which statistically manifests itself as a difference in the response distribution between the BTS group and the control group in a specific direction), would it not be useful to compare that difference to the difference between a group in which an alternative manipulation was used and the control group? I think an analysis as the one proposed by Reviewer 2 may be in line with this.

Response 3: We agree! we will conduct such analyses as exploratory additional analyses.

Comment 4: In the Abstract, at the end, it says that under ii) you want to test 'whether the effect may be explainable by an increase in expected earnings or the addition of a prediction task'. This very much sets the expectation that it is these two explanations that are being primarily considered as possible explanations of the effect. However, implicit in your text is that the primary explanation is that it is the truth incentivising interaction between the instruction and related monetary incentive that elicits the effect, and as such these would be alternative explanations. This is made more explicit in other areas of the report where you explicitly state the term 'alternative explanations' and especially 'the worry that …'. **The reader would be greatly aided in understanding the report if this was all standardised and made clearer throughout the text.**

Response 4: Thank you for pressing us on this, we have now adjusted the language throughout and hope to have addressed this worry, clearly stating that our primary hypothesis

and assumption is investigating whether the Bayesian Truth Serum is distinct from the effect of its individual parts (i.e. additional earnings and prediction tasks) We hope this has now been made adequately clear.

Example (p. 7): Our goal in these analyses is to understand if the Bayesian Truth Serum itself is distinct from an increase in earnings or the prediction task, which would bolster the claim that it should be applied more widely.

Comment 5: The last sentence of the abstract is not very clear – i.e., it is not clear how this relates to what you are testing here.

Response 5: Thank you for pointing this out, we have now rewritten this part.

Comment 6: '*While there have been significant methodological advances in psychology and cognate disciplines recently,…*' **This statement should be supported by references**, but also maybe explained further, as it is currently not sufficiently clear how it is relevant to the study at hand. In many ways, it is not necessary or useful to state this here at all?

Response 6: Thank you for this point; we now provide references to make this claim more grounded in the literature and better contextualise its importance.

Example (p. 2): While there have been significant methodological advances in psychology and cognate disciplines over the past decade (e.g. Nosek & Lakens, 2014; Nosek & Lindsay, 2018; Hales, Wesselmann, & Hilgard, 2019), there has been comparatively little work on the issue of incentivisation, i.e. the way participant responses are rewarded monetarily for their time and effort in experiments and surveys.

Comment 7: When you state that "*many papers do not report the compensation fee that was offered to research participants and the fact that these fees vary widely among the papers that do disclose them (e.g., Keith et al., 2017; Rea et al., 2020)*' **more information would be useful:** Are there actual numbers indicating prevalence available? Is the compensation here meant for the same task/time invested by participants?

Response 7: Thank you for pressing us on this. We have reported the available data from other fields in a footnote now, while pointing out that, to the best of our knowledge, no good numbers for the social sciences are available. We present the numbers from the last three years for the journal Experimental Psychology in support of our claims. Experimental Psychology is

known for its transparency and openness, and we therefore believe these numbers provide an underestimate of the identified problems.

Example (p. 3, footnote 4): Our own investigation of publications from 2019-2021 in the journal Experimental Psychology, suggests that the situation is somewhat better in psychology, perhaps because many psychological studies rely on students who participate in exchange for course credit or as part of a course requirement (30%). Among the publications that mentioned monetary compensation (43%), 31% provided no indication of the amount and only 21% expressed the amount in function of time spent.

Comment 8: '*Perhaps this is due to the null findings reported by the majority of studies that investigated the influence of financial incentives on data quality (e.g., Buhrmester et al., 2011; Crump et al., 2013; Mason & Watts, 2010; Rouse, 2015). There are, however, noteworthy exceptions indicating that increasing financial compensation can improve data quality (Ho et al., 2015; Litman et al., 2015).*' **More information here may be useful to the reader.**

Response 8: We have now added some more context to this section.

Example (p. 3, footnote 4): To the best of our knowledge, no systematic review of this question has been conducted in the context of the social sciences. However, previous work in the context of occupational research has found that a majority of studies did not report "on any aspect of the compensation system" (Clay, Berecki-Gisolf, & Collie, 2014, 111), while the results from the broader context of medicine found that "only 13.5% [of articles surveyed] mentioned financial compensation in any way, and only 11.1% listed amounts" (Klitzman, Albala, Siragusa, Nelson, & Appelbaum, 2007, 61). Our own investigation of publications from 2019-2021 in the journal Experimental Psychology, suggests that the situation is somewhat better in psychology, perhaps because many psychological studies rely on students who participate in exchange for course credit or as part of a course requirement (30%). Among the publications that mentioned monetary compensation (43%), 31% provided no indication of the amount and only 21% expressed the amount in function of time spent.

Comment 9: In the next part of the introduction (but also in a later part, where you write about incentive compatible and incentive incompatible designs), the reader receives information about participants, especially in online studies, clicking through items in surveys rather engaging with the item content in order to maximise payoff. This is contrasted with the BTS manipulation to incentivise honest answers. What is missing or what I think could be confusing to readers is the step that connects these two issues, because honest answers are not necessarily the only answers that participants may give even when they engage with the items and their content.

Response 9: Thank you for pressing us on this, we have now made this connection clearer throughout the manuscript.

Comment 10: '*When participant payments are primarily dependent on completion of the online survey, participants are likely to complete studies as quickly as possible and to complete as many of them as feasible in the time they have available in order to maximise payoffs.'* Are there any data supporting this argument? **Given the footnote, do we know how many participants fail the attention checks? Or data on differences between online and in-person studies?**

Response 10: Thank you for this comment! As far as we know there are no direct empirical tests of this claim directly, though it follows from economic theory that, unless there are other specific preferences at play, participants will want to maximise their payoffs (which they can do by either reducing the time invested per survey or to increase the numbers of surveys taken, both of which have the same effect). We have added some additional data on the online-vs-conventional sample distinction in footnote 3, though, which shows at least a higher level of attention check passes – which the authors explain by higher exposure to studies and thus better learning (and not necessarily as them being more thoughtful participants).

Example (p. 2, footnote 3): The fact that online studies include attention checks is prima facie evidence in favour of the claim that participants aim to rush through surveys in maximising their expected payoffs. About 10% of participants do not pass attention checks in MTurk studies (Barends & de Vries, 2019; Paas, Dolnicar, & Karlsson, 2018). There is evidence that MTurk samples, due to a higher exposure to studies and thus increased ability to learn, are better at attention checks than conventional student samples (Hauser & Schwarz, 2016). The fact that MTurk participants tend to be less naïve than Prolific participants might also explain why the latter fail attention checks more often (Peer, Brandimarte, Samat, & Acquisti, 2017).

Comment 11: 'The Bayesian Truth Serum works by informing participants that the survey they are about to complete makes use of an algorithm for truth-telling that has been developed by researchers at MIT and has been published in the journal Science. This algorithm will be used to assign survey answers an information score, indicating how truthful and informative the answers are. The respondents with the top-ranking information scores will receive a bonus in addition to the base pay for participation. Participants then go on to answer study items as they normally would,' **For readers who are not so familiar with the BTS manipulation, it would be useful to state more clearly whether the part of 'This algorithm will be used…' is part of the instructions given to the participants**. On a side note, the actual wording in the Schoenegger (2021) differs so it would be good to state clearly what

wording you will use in the proposed study, and also alert the reader to any differences from the Schoenegger (2021) study given that it is these results that you are aiming to replicate.

Response 11: Thank you for this comment. We now make this clearer, explaining exactly what will be shown to participants, and also explicitly give the text in Figure 3. We have also decided to change the wording back to the original 2021 instruction text to ensure that this paper can provide convincing evidence in favour of or against direct replication of the original result.

Example (p. 4-5): The Bayesian Truth Serum fundamentally works by informing participants that the survey or experiment they are about to complete makes use of an algorithm for truth-telling that has been developed by researchers at MIT and has been published in the academic journal 'Science' (see Figure 1 for specific instructions). They are told that this algorithm will be used to assign to their survey answers an information score, indicating how truthful and informative their answers are. They are also informed that the respondents with the top-ranking information scores will receive a bonus in addition to the base pay for participation. Participants then go on to answer study items as they normally would, as well as provide predictions as to the answers chosen by the total sample. See Figure 2 for an example of the prediction task needed to calculate the information scores. After the conclusion of the study and the payment of the standard participation fee, those with the highest information scores are rewarded with their additional payments (cf. also Witkowski & Parkes, 2012; Radanovic & Faltings, 2013).

Recent work by researchers at MIT that has been published in the academic journal Science has led to the development of an algorithm for detecting truth telling. In this survey we use this algorithm to determine how truthfully you answer.

We will assign an information score to your responses below which indicates how truthful and informative you are being. Once we have collected all of the responses to this survey, we will rank the survey responders by the sum of their information scores and **award a bonus of £1 to all responders in the top 1/3rd.** This bonus is paid in addition to the base pay for participating in the survey. If you would like to receive this bonus you should just answer all questions honestly.

Figure 1. Bayesian Truth Serum Text

Comment 12: 'However, participants are only told that they can earn a bonus for answering truthfully and are not informed about the specific mechanisms of the compensation scheme.' **This requires more explanation**, especially in light what instruction the participants actually receive (see a previous comment).

Response 12: Thank you for pressing us on this. We now report this in more detail.

Comment 13: 'to ensure that the results found there generalise to a new sample and effects of the Bayesian Truth Serum are as such also likely to replicate in other people's implementations.' Maybe researchers' instead of people's?

Response 13: Thanks for the suggestion, we have now changed it!

Comment 14: Will participants who took part in Schoenegger (2021) be able to take part in this study?

Response 14: We really missed this – thank you! We now clarify that those who partook in the previous study will not be able to participate in this study.

Example (p. 8): We will not recruit participants who partook in Schoenegger (2021).

REVIEWER 1:

Comment 1: 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**.

Response 1: Thank you for pointing this out. We fully agree with your assessment and we have decided to change the wording back to the original 2021 instruction text to ensure that this paper can provide convincing evidence in favour of or against replication of the original result.

Example (p. 10-11): Those in the Bayesian Truth Serum Condition and in the Additional Money Condition will receive additional payment. Those in the Bayesian Truth Serum condition will receive an introduction to the Bayesian Truth Serum based on the original one introduced by Prelec (2004). Figure 1 contains the specific wording used in this study, which is the same as used in Schoenegger (2021).

Comment 2: **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**.

Response 2: Thank you very much for outlining this possibility. While we see the potential of adding this condition, we ultimately decided not to add this condition. Allow us to outline our reasons.

First, this proposed condition is, in fact, quite similar to the one we termed 'Additional Money', where we incentivise people to produce high quality answers by paying out the same type of bonusses as in the BTS treatment. Importantly, in this condition the predictions are made after the main items (to equalise earnings per hour) so they cannot impact the decisions and any results will be able to pinpoint simply the effect of the additional compensation. The only difference between this condition and your proposed condition would be the mention of the BTS mechanism specifically, specifying exactly how these top third of participants will be identified.

Second, while you are correct that one could, in principle, only have predictions be elicited from a few 'holdout' participants, we argue that this approach is not applicable for a wide range of cases. First and foremost, having only a small subset of participants provide predictions threatens to have the predicted frequency of the full population come apart from the predicted frequency of the holdout group. This has significant consequences for the BTS mechanism as it threatens to reward the wrong types of answers (i.e. not reward the answers that would have actually been surprisingly common if all participants had provided predictions). While this approach may reduce researcher costs, we argue that doing so would be potentially detrimental to the coherence of the mechanism itself, as the predictions of a few may have significant impacts on the pay-out structure overall. Additionally, these discrepancies may accrue over time in that participants who are repeatedly exposed to an incentivisation mechanism that doesn't reward the most common answers might actually learn not to respond truthfully, which would defy the point of including the mechanism. In short, we deem the risks of this procedure substantial as long as we don't know the minimal N for the hold-out sample.

Third, we, like most researchers, are facing resource constraints that would make the adoption of this condition, even if we thought that its conclusion would add to the paper, prohibitive. We follow Lakens (https://psyarxiv.com/9d3yf/) in stating our resource constraints openly and honestly as an additional reason for why we chose not to add this condition. As such, we believe that the conditions we propose have higher expected scientific value than this one, and given

the resource constraints that we are facing, we have decided not to add this condition. We hope that our arguments have convinced you, but we are also happy to make further amendments based on your comments should you think they are needed.

Comment 3: 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**

Response 3: Thank you for highlighting this! We agree with your statement overall, in that there may be relevant heterogeneity regarding the topics of the items. However, because the seven items are so broadly distributed, including both areas that you indicate as potentially problematic as well as others that you agree are prima facie interesting to study, we argue that there are actually significant upsides to retaining the same item set. For example, doing so allows for a better replication overall of the original result, i.e. we will be able to attempt to replicate the pattern of data in which moral and non-moral topics differ in BTS susceptibility. Additionally, changing which items to pick would also raise the same worry you raise in Response 1, which means that retaining the same seven items helps keeping this study a full replication. Further, retaining the same items set allows us to have a well-justified expected effect size mechanism, which would be less straightforward with a reduced set. Additionally, as has been raised by other reviewers, we probably should not pre-select (either prior to the study or after the initial analyses) which items to include in the analysis. In order to comply with our responses to other reviewers, we will also ensure here that we do not pre-select items and will thus retain the full set of seven items.

REVIEWER 2:

Comment 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.

Response 1: Thank you for pointing this out and for giving clear-to-follow steps on improving the structure. We have now rearranged the paper as requested to aid comprehension and overall flow. We hope that it is now adequately structured but are happy to make further adjustments should you see those as necessary.

Comment 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 nonparametric 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.

Response 2: Thank you so much for pointing this out: We agree that it is not very likely that such a pattern of data may arise, but we had not properly anticipated this possibility in our previous report. On your first worry, we have now added a section in Potential Results/Implications that picks up this point and outlines its implications, i.e. that there may be ameliorative effects that may make a straightforward interpretation of our results difficult, and that potential changes in distributions may go different ways. As we see no straightforward way to statistically account for this problem, we will make sure to point this out in detail in the limitations sections to ensure that readers are informed about this and adjust our conclusions drawn from this study accordingly while making this clear throughout the manuscript.

However, it is also important to point out that our conclusions are different than the ones you propose here: In the case where we find both a significant difference between the BTS and the Control as well as between the BTS and the Prediction/Additional Money task, we do not conclude that the BTS effect is driven by the prediction task/additional money. Quite the contrary (at least in our revised manuscript), we would conclude that there is something distinct to the Bayesian Truth Serum that is neither capture by an increase in earnings nor by the prediction task. To arrive at such a conclusion the pattern of data as outlined by you above, it would have to be such that the BTS differs from the Control, but that we fail to find a statistically significant difference compared to, for example, the prediction task. Now, when we fail to be able to distinguish the prediction task from the BTS, we will conclude that the effect of the BTS is not distinct from that of prediction (and that, in effect, the BTS may be driven by it, though additional exploratory analyses may have to be conducted). We understand that the previous version of the manuscript really did not make this clear at all and we take full responsibility for this. We hope that with this now being properly clarified throughout the manuscript, that your worry applies less than first feared.

On your second worry, we agree that the sequencing of analyses was problematic as outlined in our original report. We now have our analysis such that we do not filter conditions by significant comparisons in the main comparisons like we had previously proposed but instead analyse all sets of comparisons at the same time irrespective of the other results.

Example (p. 17-18): In more general terms, we consider the following potential patterns of results. First, there is the pattern of results where we find significant differences between the No Incentive Condition and the Bayesian Truth Serum Condition, suggesting a successful replication, while also finding significant differences between the Bayesian Truth Serum Condition and the Additional Money Condition as well as the Prediction Condition. In this case, the evidence would point towards a unique effect of the Bayesian Truth Serum in the context of Likert-scale items and would provide a solid basis for the adoption of this mechanism in psychology and experimental philosophy.

A second pattern of results is one where we do find a difference between the No Incentive Condition and the Bayesian Truth Serum Condition but fail to find a significant difference between the Bayesian Truth Serum Condition and one (or both) of the other conditions. In this case, while we do provide evidence in favour of a replication of the effect of the Bayesian Truth Serum, we provide mixed or inconclusive evidence in favour of the distinct nature of the Bayesian Truth Serum. It might be that we find evidence that the BTS effect might be driven by the addition of the prediction task or the increase in compensation. In this case, we would not make a recommendation for an adoption of this mechanism in psychology and experimental philosophy but will provide further avenues for research.

A third pattern of results is one where we fail to provide evidence in favour of a replication. In this case we clearly would not make a recommendation for the adoption of this mechanism. However, there are also patterns of data in which we fail to provide evidence in favour of a replication, while at the same time finding a difference between the Bayesian Truth Serum Condition and one of the other two conditions. However unlikely that may be, this is a potential pattern of results. In these cases, we could conclude that while we fail to provide

evidence in favour of a replication, we find that, for example, adding prediction tasks does affect the answer distributions more than the combined effect of prediction plus additional compensation plus Bayesian Truth Serum framing. In other words, one may think that the combined Bayesian Truth Serum may reverse or ameliorate some of the effects of its constituent parts. This will again open up new research questions that would focus on this specifically.

Comment 3: 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 an **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.**

Response 3: Thank you very much for your comment. While we agree that the proposed analysis plan does not allow for the overall identification of effects, we have decided not to follow your recommendation because the alternative analysis you propose is essentially concerned with changes in the mean response, while previous work on the BTS has focussed on changes in the (entire) response distribution instead (e.g., Frank et al. 2017; Weaver & Prelec, 2013). As a result, evidence that the BTS affects the mean response is currently lacking. Indeed, when re-analysing the original data from Schoenegger (2021), we found no statistically significant mean differences, which further reinforced our plan to focus on distributional changes. Unfortunately, we do not know of an analysis that is akin to the one you proposed that is concerned with distributions rather than means, which is why we decided to retain our original analysis plan.

Comment 4: 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.)

Response 4: Thanks! We have now updated our discussion of the literature to more fully draw on the work in economics to properly contextualise our work.

Example (p. 3): The main exception to this claim is the field of economics, where incentive compatible research designs (both involving areas with objective as well as subjective

data) have both been discussed and applied widely (e.g. Hertwig & Ortmann, 2001; Offerman, Sonnemams, Van De Kuilen, & Wakker, 2009; Schlag, Tremewan, Van der Weele, 2015; Baillon, 2017).