- In the abstract – the "We fail to replicate the effect of the BTS on response behaviour and are as such unable to recommend wide adoption of the BTS on the basis of these data." 1. 'on the basis of these data' is redundant? 2. Could be better phrased, i.e. more specificly, like in the discussion section (p 17 of stage 2 manuscript with tracked changes)?

  "*Accordingly, we are unable to make a recommendation for the adoption of the BTS mechanism in social science fields that rely heavily on Likert-scale items reporting subjective data as we have studied in this context.* "

  This is also more in line with what was written in the stage 1 report (p.16 of stage 1 manuscript, v2).

  " *This will mean that we will not be able to recommend the adoption of the BTS in the context of psychology and experimental philosophy in the form studied here.*"

Response: Thank you for pointing this out. It is changed accordingly.

- Figure 4 requires a legend such that the acronyms/labels of the different plots are understandable.

Response: Added!

- p.15 "For all analyses below (and in Appendix B), we use the pre-registered adjusted significance threshold of .007 and designate it with '***'. Effects with p-values greater than this (p < .01 or '**', p < .05 or '*', and p >. 05 without stars) are interpreted as non-significant."

  In the stage 1 report, the justification for the significance level of .007 was given. It is not clear why the other levels are mentioned or highlighted in the results, given that they do not relate to any inferences in the present study.

Response: We had included these for full presentation of results in a clear and space-efficient manner. As we realize that these might make readers infer significance where there is none according to our pre-registered criteria, we have now removed all indications except for p < .007.

- Related to the above, the tables could be more informative, i.e., the exact p-values could be given from the analysis in question, with those smaller than the significance level (.007) highlighted.

Response: We fear that including the p-values may make the table much more crowded and would lead to a similar concern as indicated in response to the previous comment: it may make readers infer significance (e.g., when encountering a $p < .05$) where they shouldn't based on our pre-defined criteria. Therefore we prefer to not mention the p-values (seeing that they can also be determined based on the test statistics and degrees of freedom) and include the effect sizes instead.

- Please ensure that the comparison between the treatments clearly labelled in terms of being an additional non-preregistered analysis.

Response: Thank you so very much, we absolutely missed this, and you are correct on this. We've now labelled them as such both in the text and in the tables.

- For the comparisons between the treatments, it is not clear whether the adjusted significance level of .007 is adequate. This should be explained more clearly.

Response: Response: Apologies if this was unclear, but we simply use the same adjustment as we conduct the same number of tests (7) per comparison. We've now made this clearer.