**Taking A Closer Look At The Bayesian Truth Serum: A Registered Report**

*Schoenegger Philipp[[1]](#footnote-1)\* & Verheyen Steven[[2]](#footnote-2)*

**Abstract**

Over the past decade, psychology and its cognate disciplines have undergone substantial scientific reform, ranging from advances in statistical methodology to significant changes in academic norms. One aspect of experimental design that has received comparatively little attention is incentivisation, i.e. the way that participants are rewarded and incentivised monetarily for their participation in experiments and surveys. While incentive compatible designs are the norm in disciplines like economics, the majority of studies in psychology and experimental philosophy are constructed such that individuals’ incentives to maximise their payoffs in many cases stand opposed to their incentives to state their true preferences honestly. This is in part because the subject matter is often self-report data about subjective topics and the sample is drawn from online platforms like Prolific or MTurk where participants are out to make a quick buck. One mechanism that allows for the introduction of an incentive-compatible design in such circumstances is the Bayesian Truth Serum (Prelec, 2004), which rewards participants based on how surprisingly common their answer are. Recently, Schoenegger (2021) applied this mechanism in the context of Likert-scale self-reports, finding that the introduction of this mechanism significantly altered response behaviour. In this registered report, we further investigate this mechanism by (i) directly replicating the previous result and (ii) testing whether the effect is best explained by parts of the mechanism like the increase in expected earnings or the addition of a prediction task. We hope that this paper can provide further evidence for the adequacy of the Bayesian Truth Serum as an incentivisation mechanism in a variety of social science fields like psychology and experimental philosophy.

Keywords: *Incentivisation, Bayesian Truth Serum, Methods, Open Science*

**Introduction**

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. The central worry expressed in this paper is that this failing to take the issue of incentivisation seriously can negatively affect the quality of collected data, particularly in a context where data is increasingly crowdsourced from an online population that regards participation in online research as their main source of income (Eyal et al., 2021). When participant payments are primarily dependent on completion of an online survey or experiment, participants are likely to complete studies as quickly as possible and to complete as many of them as is feasible in the time they have available in order to maximise their personal payoffs.[[3]](#footnote-3) However, as researchers, we want participants to take their time with the study items and respond carefully and truthfully. That is, we want to collect data from participants who took their time to properly read the instructions, engaged with the material, and revealed their honest preferences in self-report measures or behaviour.

Most social sciences have so far failed to systematically engage with the question of how to properly incentivise research participants (beyond the status quo that posits that simply paying participants a completion fee is sufficient). This is reflected in the casual observation that many papers do not report the monetary compensation fee that was offered to their research participants[[4]](#footnote-4) 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). Perhaps this neglect of incentivisation 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), with some noteworthy exceptions (Ho et al., 2015; Litman et al., 2015). All in all, however, there has been a concrete lack of engagement with incentivisation mechanisms across much of the social sciences.

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). In a recent paper, Schoenegger (2021) draws on this literature and presents this incentivisation challenge in detail, proposing the adoption of a potential solution applicable to experimental philosophy as well as related disciplines like psychology: Their suggestion is to use the Bayesian Truth Serum (BTS) first introduced by Prelec (2004) to improve data quality and to allow for incentive compatibility in several academic fields where this is not currently the norm.

The Bayesian Truth Serum (Prelec, 2004) is an incentivisation mechanism primarily for research where the subject matter is subjective (i.e. where researchers cannot score participant answers as ‘true’ or ‘false’), as is the case for much of the research conducted across the social sciences. According to this proposal, researchers would apply a post-hoc incentivisation scheme to address worries regarding data quality by rewarding participants financially for answering truthfully. This represents an incentive compatible mechanism aligning participants’ profit maximising motives with their motives to state their honest views and preferences. As such, in an incentive compatible design, participants can maximise their expected payoff by answering truthfully, while an incentive incompatible design sees these two forces come apart; participants may eschew answering honestly to maximise profits. The latter is problematic for scientific research as the data might become invalid and conclusions drawn from them potentially wrong (Weaver & Prelec, 2013).

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).

Figure 1. Bayesian Truth Serum Text

Participants are awarded the bonus, both on the basis of how well their predictions fit the actual distribution of answers and how surprisingly common their own answers are. However, participants are only told that they can earn a bonus for answering truthfully and are not informed about the specific mechanisms of the post-hoc compensation scheme. The central criterion of surprisingly common answers derives its theoretical justification from the Bayesian claim that the “highest prediction of the frequency of a given opinion […] should come from individuals who hold that opinion” (Prelec, 2004, p. 462). As such, rewarding surprisingly common answers is akin to rewarding honest answers (Prelec, 2004).

Figure 2. Example Bayesian Truth Serum Prediction Item

The Bayesian Truth Serum has already been validated in a large-scale online study setting on MTurk (Frank et al., 2017) and has already been applied in a variety of contexts, including in marketing (Howie et al., 2010), metascience (John et al., 2012), criminology (Loughran et al., 2014), and economics (Zhou et al., 2019). As outlined above, the Bayesian Truth Serum is a natural incentivisation mechanism for research in psychology and experimental philosophy. Because the subject matter is inherently subjective, one cannot otherwise ascertain which answers are honest or correct. Schoenegger (2021) report the application of the Bayesian Truth Serum on a number of questions drawn from papers published within the last ten years either in *Philosophical Psychology* or in *The Review of Philosophy and Psychology*. In a Prolific sample, they show that “regular” response patterns differ significantly from responses that have been incentivised by the Bayesian Truth Serum and propose that the mechanism be adopted by experimental philosophers and psychologists more widely.

However, while there has been significant work on the Bayesian Truth Serum and its underlying mechanisms (e.g. Weaver & Prelec, 2013; Frank et al., 2017), there remain a number of questions regarding its application that previous work has not yet addressed. In this paper our central aim is to (a) directly replicate the most recent results by Schoenegger (2021) to ensure that the results found there robustly generalise to a new sample and that the effects of the Bayesian Truth Serum are as such also likely to replicate in other researchers’ work. Further, (b) we aim to investigate if the Bayesian Truth Serum is distinct from its constituent parts, for example by looking at whether increased monetary compensation matching the expected earnings of the participants incentivised by the Bayesian Truth Serum could explain the shift in responses, as one might wonder whether any given effect of the Bayesian Truth Serum may be primarily due to increased expected earnings. Additionally, (c) we also investigate whether the prediction process itself could explain the change in response patterns by including a condition where participants are not incentivised by the Bayesian Truth Serum while still providing the same predictions, as one plausible explanation for any found effect of the Bayesian Truth Serum may simply be that the prediction task induces reflection that affects participants’ answers such as to explain the previous results. 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.

**Methods**

In order to determine the sample size needed, we conducted an a priori power analysis. 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.

With the above effect size of Cramer’s V=.1 (φ=.245) and assuming an alpha level of .007 (correcting for multiple hypothesis testing according to the Bonferroni-method based on the seven tests we will conduct for each comparison), and a power of .80, as well as df=6, the projected total sample size needed for each pairwise comparison is 333. As we will have four conditions that are each evaluated pairwise with each other, we estimate that we need at least 666 participants. In order to adjust for the exclusion rate (at around 5% in Schoenegger, 2021), we will recruit a total of 700 participants – 175 for each condition.[[5]](#footnote-6),[[6]](#footnote-7) We will not recruit participants who partook in Schoenegger (2021).

We estimate the costs of this to be the following: As we will recruit these participants from Prolific with a base pay of £1 for a participation of around 10 minutes, the base costs for participant payments will be £980. We will then incentivise participants in the Bayesian Truth Serum Condition according to the mechanism. Following the original paper, the top third, i.e. about 55 of the roughly 166 participants total, will receive an additional £1. Factoring in the Prolific charge for bonus payments, this will add another £77. We will also need the same budget increase for the Additional Money Condition where we control for expected earnings, leaving us with an anticipated cost of £1134 for this research project, which is in line with our actual budgetary constraints.

Participants from the UK will be recruited via Prolific and then randomly assigned to one of the four conditions, see Figure 3. All participants will be presented with the same list of items as used by Schoenegger (2021) that utilise Likert-scales commonly used in psychology and experimental philosophy. Specifically, we will include the items on attributions of knowledge-how in conditions of luck (Carter et al., 2019) – item 1, modesty (Weaver et al., 2017) – item 2, freedom of choice in situations of nudging (Hagman et al., 2015) – item 3, the moral permissibility of torture (Spino & Cummins 2014) – item 4, the correspondence theory of truth (Barnard & Ulatowski 2013) – item 5, moral responsibility (De Brigard & Brady 2013) – item 6, and determinism (Nadelhoffer et al., 2020) – item 7. All items will be accompanied by a 7-point Likert scale ranging from 1= ”Strongly disagree” to 7= ”Strongly agree”.[[7]](#footnote-8) See the Appendix for all seven items. The items will be presented in a random order to participants. Participants will be randomly selected into one of four conditions. Below we outline our overall experimental procedure in Figure 3.

Diagram

Description automatically generated

Figure 3. Experimental Outline.

Those in the Bayesian Truth Serum Condition and in the Prediction Condition will be asked to also provide predictions as to the underlying distribution of answers on the same page where they provide their own response (see Figure 2 for an illustration). Specifically, they will have to provide the frequency of every of the seven answers (1 through 7) to each item, with the constraints that each estimate cannot by itself be smaller than ‘1’ and they all have to sum to ‘100’. Participants in the No Incentive and Additional Money Conditions will also be asked to provide predictions (to hold constant earnings per time), though they will only make these predictions once they have provided their own answers to all seven items and have moved on to the next page.

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).

The participants in the other conditions will simply be told how they will be compensated for their participation, with the exception of those in the Additional Money Condition, who will be presented with a formulation similar to the BTS explaining that the top third of quality responses will receive a bonus. The specific wording will be: “We will award a quality score to your responses below. Once we have collected all the responses to this survey, we will rank the survey responders by the sum of their quality 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 this survey”.

**Hypotheses**

We have three distinct goals. Specifically, we investigate (i) the replicability of the original finding, (ii) analyse any effect of the Bayesian Truth Serum is distinct from an increase in expected earnings that accompanies the Bayesian Truth Serum, and (iii) analyse whether any effect of the Bayesian Truth Serum is distinct from the addition of the prediction task itself. Below we outline these goals in more detail and state our hypotheses clearly.

(i) **Replication.** First, we would like to attempt to directly replicate the finding from Schoenegger (2021), as it is a recent finding that uses question types commonly used across the social sciences (i.e. Likert-scale self-report items), and as such has a high potential applicability in fields where incentivisation mechanisms are largely absent (i.e. psychology and experimental philosophy). We take replication to be crucial for a cumulative science as a single finding ought not to be taken as sufficient evidence for potentially wide-ranging costly reforms like the introduction of an incentivisation method that significantly increases both time and monetary costs of research. To bolster the evidentiary basis for the claim that social scientists ought to adopt the Bayesian Truth Serum in the context of psychology and experimental philosophy, one ought to be reasonably confident that the mechanism has a measurable and replicable impact on responses to items commonly used in these fields.

In order to test whether one can be confident in the action recommendation Schoenegger (2021) outlined in the context of experimental philosophy, we will therefore see whether we can replicate their finding that applying the Bayesian Truth Serum yields different response patterns compared to the default practice of paying participants for study completion. We further add to the standard control condition (No Incentive Condition) a prediction task directly after the main study items to hold constant earnings per hour across all conditions. This should further bolster our confidence in any given result. Importantly though, this added prediction task cannot influence the main responses as participants first have to complete all main study items. That is, the prediction task does not accompany individual study items as is the case in the Bayesian Truth Serum Condition illustrated in Figure 1.

(ii) **Expected Participant Earnings.** One concern related to the previous instantiations and validations of the Bayesian Truth Serum is that it is quite plausible that an observed change in response distributions might be due to a change in expected earnings (before uncertainty as to its allocation is resolved), specifically the bonuses awarded to the top third of participants in the Bayesian Truth Serum condition. After all, participants that are incentivised with the Bayesian Truth Serum receive standard participation compensation, as well as additional monetary rewards based on the honesty of their answers. This would make the Bayesian Truth Serum itself not distinct from simply raising compensation levels. In order to test whether the Bayesian Truth Serum is indeed distinct from simply increasing participant payments, we include a condition where we adjust the expected earnings of an otherwise standard control condition to match that of the Bayesian Truth Serum treatment: the Additional Money Condition. In order to test whether this additional monetary reward is driving the potential change of response distributions in the Bayesian Truth Serum Condition, we test whether answer distributions differ from those produced by the Bayesian Truth Serum treatment. The additional reward is also provided in the form of a bonus of the same size to a third of participants to keep constant the probabilistic nature of the additional compensation. As before, we also include a post-study prediction task to hold constant time spent on the study and to properly equalise expected earnings per time.[[8]](#footnote-9)

(iii) **Prediction Task.** Lastly, it may also be that any effect established by the Bayesian Truth Serum might be due to participants having to give predictions while those in the control conditions typically do not have to complete a similar task. In other words, it may be that the empirical evidence speaking in favour of an effect does not stem from the Bayesian Truth Serum instructions, but instead from the fact that those who are in the treatment conditions also have to provide predictions that impact their own responses. This would make the Bayesian Truth Serum not distinct from simply adding a prediction task. To investigate this, we test whether simply adding a prediction task has a similar effect as the Bayesian Truth Serum. In the Prediction Condition, participants will therefore answer the main study items and the accompanying prediction tasks simultaneously (as they would if incentivised by the Bayesian Truth Serum). The difference with the Bayesian Truth Serum condition is that participants in the Prediction Condition do not have the chance to obtain bonus payments. This allows us to identify whether the Bayesian Truth Serum is indeed distinct from simply adding a prediction task.

**Analysis Plan**

In our primary analysis, we will follow the same procedure as Schoenegger (2021). Specifically, we will conduct a Pearson’s χ2 Goodness-of-Fit test for pairwise distribution comparisons. We made the choice to focus on a change in distribution because the vast majority of previous papers has used this (or a very similar) test. As such, there is a lot of evidence that the BTS impacts response distributions (Frank et al. 2017; Weaver & Prelec, 2013).

As the main analyses will be compromised of seven individual comparisons, one for each item, we will adjust our p-values based on the Bonferroni method to an alpha of .007 in order to correct for multiple hypothesis testing. Given our power analysis outlined above with our expected effect size of V=.1, we will take non-significant effects as a rejection of the null up to this effect size, and we remain agnostic as to effects smaller than V=.1. Given the sample size we will collect, this allows us to detect an effect at least as large as the expected effect of .1 with a probability of .80. For all tests relating to all three hypotheses, we will also report χ2 Tests of Homogeneity as well as Kolmogorov-Smirnov tests in the appendix to show sensitivity of results to the choice of statistical test. We will also conduct exploratory analyses comparing the difference of the main analyses to the difference between the Control Condition and one of the additional conditions.

For hypothesis (i), aiming to provide a direct replication of the original effect, we will code the response pattern for the No Incentive Condition as the expected distribution and the response pattern for the Bayesian Truth Serum Condition as the observed distribution. Then we conduct the Pearson’s χ2 Goodness-of-Fit test for pairwise distribution comparisons. We will treat a pattern of data that shows significant changes in response patterns in at least four out of seven items as strong evidence, as the original paper reported significant differences for four items (at the *p*<.001 level even though it did not explicitly adjust for multiple comparisons). We will treat a pattern that shows significant differences in response patterns in between one and three out of seven items as weak evidence. We will treat patterns that show no significant differences as evidence in favour of a null effect of up to V=.1, and as such a failure to replicate.

For hypothesis (ii), aiming to identify whether the Bayesian Truth Serum is distinct from a mere increase in expected earnings, we will code the response pattern for the Bayesian Truth Serum Condition as the expected distribution and the response pattern for the Additional Money Condition as the observed distribution. Then we conduct the Pearson’s χ2 Goodness-of-Fit test for pairwise distribution comparisons. As in (i), we will treat response patterns with at least four significant differences as strong evidence for a difference and patterns with between one and three differences as weak evidence for a difference. We will treat patterns that show no significant differences as evidence in favour of a null effect of up to V=.1 and we suspend judgement for effects smaller than that.

The analysis for hypothesis (iii) will be identical to that for hypothesis (ii) with the response pattern for the Prediction Condition instead of the response pattern for the Additional Money Condition as the observed distribution.

**Potential Results/Implications**

Below we outline the potential results and implications for each of the three hypotheses.

(i) Either our data provide strong support for the replicability of the effect, or the data provide weak support, or the data do not provide support for the claim that this effect is replicable. In the first two cases, we will describe the replication as being successful if either strong or weak evidence for it is present, but properly describe the evidence for the replication as either ‘strong’ or ‘weak’ depending on the criteria set out above. If we do not find statistically significant differences, we will understand this as a failure to provide evidence in favour of a successful replication. 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.

(ii) In investigating the role of expected earnings, we will conduct the same tests as in (i). In analysing the difference in distributions between the Additional Money Condition and the Bayesian Truth Serum Condition (which is coded as the expected distribution), we take the same approach as in (i), with four or more differences specifying strong evidence in favour of a difference, between one and three differences specifying weak evidence in favour of a difference, and no differences as evidence in favour of a null effect up to V=.1. In finding (weak or strong) evidence for differences, we will have provided (weak or strong) evidence in favour of an effect specific to the Bayesian Truth Serum. In other words, if the answer distributions in the Bayesian Truth Serum Condition differ significantly from those in the Additional Money Condition, then we take this to provide evidence in favour of the BTS being, at least in part, responsible for the main differences compared to the No Incentive Condition. However, if we fail to find a significant effect, we take this as a failure to provide evidence in favour of the claim that the effect of the Bayesian Truth Serum is distinct from that of additional compensation.

(iii) Lastly, in investigating the role of the prediction task, we will conduct the same tests as in (i). In analysing the difference in distributions between the Prediction Condition and the Bayesian Truth Serum Condition (which is coded as the expected distribution), we take the same approach as in (i), with four or more differences specifying strong evidence in favour of a difference, between one and three differences specifying weak evidence in favour of a difference, and no differences as evidence in favour of a null effect up to V=.1. As above, in finding (weak or strong) evidence for differences, we will have provided (weak or strong) evidence in favour of an effect specific to the Bayesian Truth Serum. However, if we fail to find a significant effect, we take this as a failure to provide evidence in favour of the claim that the effect of the Bayesian Truth Serum is distinct from that of the addition of the prediction task.

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.

**References**

Baillon, A. (2017). Bayesian Markets to Elicit Private Information. *Proceedings of the National Academy of Sciences*, *114*(30), 7958-7962.

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences, 143*, 84-89.

Barnard, R., & Ulatowski, J. (2013). Truth, Correspondence, and Gender. *Review of Philosophy and Psychology*, *4*(4), 621-638.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data? *Perspectives on Psychological Science, 6*, 3–5.

Carter, J. A., Pritchard, D., & Shepherd, J. (2019). Knowledge-How, Understanding-Why and Epistemic Luck: an Experimental Study. *Review of Philosophy and Psychology*, *10*(4), 701-734.

Clay, F. J., Berecki-Gisolf, J., & Collie, A. (2014). How well do we Report on Compensation Systems in Studies of Return to Work: A Systematic Review. *Journal of Occupational Rehabilitation*, *24*(1), 111-124.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PloS one, 8*(3), e57410.

De Brigard, F., & Brady, W. J. (2013). The Effect of what we think may happen on our Judgments of Responsibility. *Review of Philosophy and Psychology*, *4*(2), 259-269.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data Quality of Platforms and Panels for Online Behavioral Research. *Behavior Research Methods*, 1-20.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical Power Analyses using G\* Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

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.

Hagman, W., Andersson, D., Västfjäll, D., & Tinghög, G. (2015). Public Views on Policies Involving Nudges. *Review of Philosophy and Psychology*, *6*(3), 439-453.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving Psychological Science through Transparency and Openness: An Overview. *Perspectives on Behavior Science*, *42*(1), 13-31.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk Participants Perform better on Online Attention Checks than do Subject Pool Participants. *Behavior Research Methods*, *48*(1), 400-407.

Hertwig, R., and A. Ortmann. 2001. Experimental Practices in Economics: A Methodological Challenge for Psychologists? *Behavioral and Brain Sciences 24* (3): 383–403.

Ho, C. J., Slivkins, A., Suri, S., & Vaughan, J. W. (2015, May). Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 419-429).

Howie, P. J., Wang, Y., & Tsai, J. (2011). Predicting New Product Adoption Using Bayesian Truth Serum. *Journal of Medical Marketing*, *11*(1), 6-16.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. *Psychological Science*, *23*(5), 524-532.

Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems Perspective of Amazon Mechanical Turk for Organizational Research: Review and Recommendations. *Frontiers in Psychology, 8*: 1359.

Klitzman, R., Albala, I., Siragusa, J., Nelson, K. N., & Appelbaum, P. S. (2007). The Reporting of Monetary Compensation in Research Articles. *Journal of Empirical Research on Human Research Ethics, 2*(4), 61-67.

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The Relationship between Motivation, Monetary Compensation, and Data Quality among US-and India-based Workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519-528.

Loughran, T. A., Paternoster, R., & Thomas, K. J. (2014). Incentivizing Responses to Self-Report Questions in Perceptual Deterrence studies: An investigation of the Validity of Deterrence theory using Bayesian Truth Serum. *Journal of Quantitative Criminology*, *30*(4), 677-707.

Mason, W., & Watts, D. J. (2009, June). Financial Incentives and the Performance of Crowds. In Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 77-85).

Nadelhoffer, T., Yin, S., & Graves, R. (2020). Folk Intuitions and the Conditional Ability to do Otherwise. *Philosophical Psychology*, *33*(7), 968-996.

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*(3), 137-141.

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the Norm in Psychological Science. *APS Observer*, *31*(3).

Offerman, T., Sonnemans, J., Van de Kuilen, G., & Wakker, P. P. (2009). A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Rttitudes. *The Review of Economic Studies*, *76*(4), 1461-1489.

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing, 35*(2), 258-269.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioural research. *Journal of Experimental Social Psychology, 70*, 153-163.

Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science*, *306*(5695), 462-466.

Radanovic, G., & Faltings, B. (2013). *A Robust Bayesian Truth Serum for Non-Binary Signals*. In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13) (no. CONF, pp. 833-839).

Rea S. C., Kleeman, H., Zhu, Q., Gilbert, B., & Yue, C. (2020). Crowdsourcing as a Tool for Research: Methodological, Fair, and Political Considerations. *Bulletin of Science, Technology & Society, 40*(3-4), 40-53.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior, 43*, 304-307.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A Penny for your Thoughts: A Survey of Methods for Eliciting Beliefs. *Experimental Economics*, *18*(3), 457-490.

Schoenegger, P. (forthcoming). Experimental Philosophy and the Incentivisation  
Challenge: A proposed Application of the Bayesian Truth Serum

Spino, J., & Cummins, D. D. (2014). The Ticking Time Bomb: When the use of Torture is and is not Endorsed. *Review of Philosophy and Psychology*, *5*(4), 543-563.

Weaver, R., & Prelec, D. (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research 50*(3): 289–302.

Weaver, S., Doucet, M., & Turri, J. (2017). It’s What’s on the Inside that Counts... Or is It? Virtue and the Psychological Criteria of Modesty. *Review of Philosophy and Psychology*, *8*(3), 653-669.

Witkowski, J., & Parkes, D. C. (2012). *A Robust Bayesian Truth Serum for Small Populations*. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI’12).

Zhou, F., Page, L., Perrons, R. K., Zheng, Z., & Washington, S. (2019). Long-Term Forecasts for Energy Commodities Price: What the Experts Think. *Energy Economics*, *84*, 104484.

**Appendix**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Question | Hypothesis | Sampling Plan | Analysis  Plan | Rationale for  deciding the  sensitivity of  the test for  confirming or  disconfirming  the hypothesis | Interpretation  given different outcomes | Theory that could  be shown wrong  by the outcomes |
| Does the  result of  Schoenegger  (2021) replicate? | NH: There is no  difference in res-  ponse distribut-  ions between the  Control  and the BTS  conditions | Chi-squared Goodness-of-Fit (φ=.245, alpha= .007, power=.8,  df=6) 🡪 n=333  (for each pair- wise comparison) - see Methods  section in paper | Pearson’s χ2  Goodness- of-Fit test for  pairwise  distribution comparisons - see Analysis Plan in paper | Small effect  size of V=.1  that is consis-  tent with avg.  previous effect  at V=.117 and  smallest prev.  significant eff.  at V=.101 | Significant differ- ences in at least 4 items will be  taken as strong evidence, 1-3 differences will be  weak evidence  - see Potential  Results in paper | Failure to replicate  would indicate that  the BTS does not in  fact reliable shift  answer distributions |
| Does increasing expected earnings account for the effect of the BTS? | NH: There is no  difference in res-  ponse distribut-  ions between the  Additional Money  and the BTS  condition | See above | See above | See above | See above | Showing that the effect of the BTS is primarily  explained by increasing expected earnings  provides evidence  against the claim that  standard BTS texts drive the effect |
| Does the predict-  ion task account  for the effect of  the BTS? | NH: There is no  difference in res-  ponse distribut-  ions between the  Prediction  and the BTS  condition | See above | See above | See above | See above | Showing that the effect of the BTS is primarily  explained by adding  a prediction task  provides evidence  against the claim that  standard BTS texts drive the effect |

**Study Design Template**

1. University of St Andrews, School of Economics & Finance; School of Philosophical, Anthropological and Film Studies, ps234@st-andrews.ac.uk [↑](#footnote-ref-1)
2. Erasmus University Rotterdam, Department of Psychology, Education and Child Studies, verheyen@essb.eur.nl [↑](#footnote-ref-2)
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). [↑](#footnote-ref-3)
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. [↑](#footnote-ref-4)
5. This research has received ethics approval from the University of St Andrews (SA15351). [↑](#footnote-ref-6)
6. All data will be made freely available upon collection. [↑](#footnote-ref-7)
7. This deviates from Schoenegger (2021) in that they had one 5-point Likert scale as the original item had a 5-point Likert scale. To ensure more consistency across all items, we chose to also use a 7-point Likert scale for this item. [↑](#footnote-ref-8)
8. This condition differs from previous work by Weaver and Prelec (2013) who study how “implement[ing] BTS without explaining the basis of the payments and without asking people to answer […] honestly” (Weaver & Prelec, 2013, 290) impacts choices, finding that in their sample of 27 participants, truth-telling incentives remain compelling. In our work, we explicitly state the additional monetary compensation upfront and tell participants that we will rank their answers by quality. This helps us more directly identify the effect of the BTS specifically as opposed to simply paying participants better for their answers. [↑](#footnote-ref-9)