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

Schoenegger Philipp[1]* & Verheyen Steven[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 many 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 (BTS; Prelec, 2004), which rewards participants based on how surprisingly common their answers 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) attempting to directly replicate the previous result and (ii) analysing if the Bayesian Truth Serum's effect is distinct from the effects of its constituent parts (increase in expected earnings and addition of prediction tasks). We fail to find significant differences in response behaviour between participants who were simply paid for completing the study and participants who were incentivized with the BTS. Per our pre-registration, we regard this as evidence in favour of a null effect of up to V=.1 and a failure to replicate, but reserve judgment as to whether or not the BTS mechanism should be adopted in social science fields that rely heavily on Likert-scale items reporting subjective data, seeing that smaller effect sizes might still be of practical interest and results may differ for items different from the ones we studied. Further, we provide weak evidence that the prediction task itself influences response distributions and that this task's effect is distinct from an increase in expected earnings, suggesting a complex interaction between the BTS' constituent parts and its truth telling instructions.

[1] University of St Andrews, School of Economics & Finance; School of Philosophical, Anthropological and Film Studies, ps234@st-andrews.ac.uk
[2] Erasmus University Rotterdam, Department of Psychology, Education and Child Studies, verheyen@essb.eur.nl

Keywords: *Incentivisation, Bayesian Truth Serum, Methods, Rewards, Bonus*

Introduction[3]

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 participants 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 to maximise their personal payoffs.[4] However, as researchers, we want participants to take their time with study items and respond

---

[4] 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; Kees, Berry, Burton, & Sheehan, 2017). Specifically, Kees et al. find that 90.8% of MTurk participants passed an instructional manipulation check, while only 64.3% of undergraduate participants did (Kees, Berry, Burton, & Sheehan, 2017, p. 149). In Hauser and Schwarz (2016, p. 403), these percentages equalled 95% and 39%, respectively. The result held up even with a novel instructional manipulation check, where 25.5% of MTurkers passed, compared to only 2.2% of undergraduate participants (Hauser & Schwarz, 2016, p. 405). 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).

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 of simply paying participants a completion fee). This is reflected in the casual observation that many papers do not report the monetary compensation fee that was offered to their research participants[5] 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,

---

[5] 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, p. 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, p. 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.

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 that claims to reward 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. 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 their 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

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 in these fields is inherently subjective, one cannot otherwise ascertain which answers are honest or correct. Schoenegger (2021) reports the application of the Bayesian Truth Serum on several 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, or (c) whether the addition of 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. These two options are plausible because one 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 or that it is simply an effect of higher compensation. 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.

Hypotheses

In this paper, we have three distinct goals and hypotheses. Specifically, we investigate (i) the replicability of the original finding, (ii) analyse whether 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 participant 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)[6] 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 2.

(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 via 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

---

[6] Though note that participants in the 'No Incentive Condition' still receive a base pay of £1.15, but they do not receive any additional incentives.

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.[7]

(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 BTS instructions, but instead from the fact that those who are in the treatment conditions also have to provide predictions that may 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 BTS). 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.

---

[7] 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.

Methods

We conducted an a priori power analysis and 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.[8] Further, as the standard 'small effect size' for Cramer's V is conventionally put at V=.1 and to be conservative, we chose 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. With four conditions that are each evaluated pairwise with each other, we needed at least 666 participants. In order to adjust for the exclusion rate (at around 5% in Schoenegger, 2021), we planned to recruit a total of 700 participants – 175 for each condition.[9] We did not recruit participants who partook in Schoenegger (2021).

Participants from the UK were recruited via Prolific and were paid £1.15[10] for participation. They were then randomly assigned to one of the four conditions,

---

[8] Note also that the smallest significant effect from the previous study was V=.101

[9] This research has received ethics approval from the University of St Andrews (SA15351).

[10] The base pay was higher than pre-registered to account for the increase in minimum payment that Prolific instituted between the submission of our pre-registration and running of this study.

see Figure 3. All participants were 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 included 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 were accompanied by a 7-point Likert scale ranging from 1= "Strongly disagree" to 7= "Strongly agree".[11] See Appendix A for all seven items. The items were presented in a random order to participants.
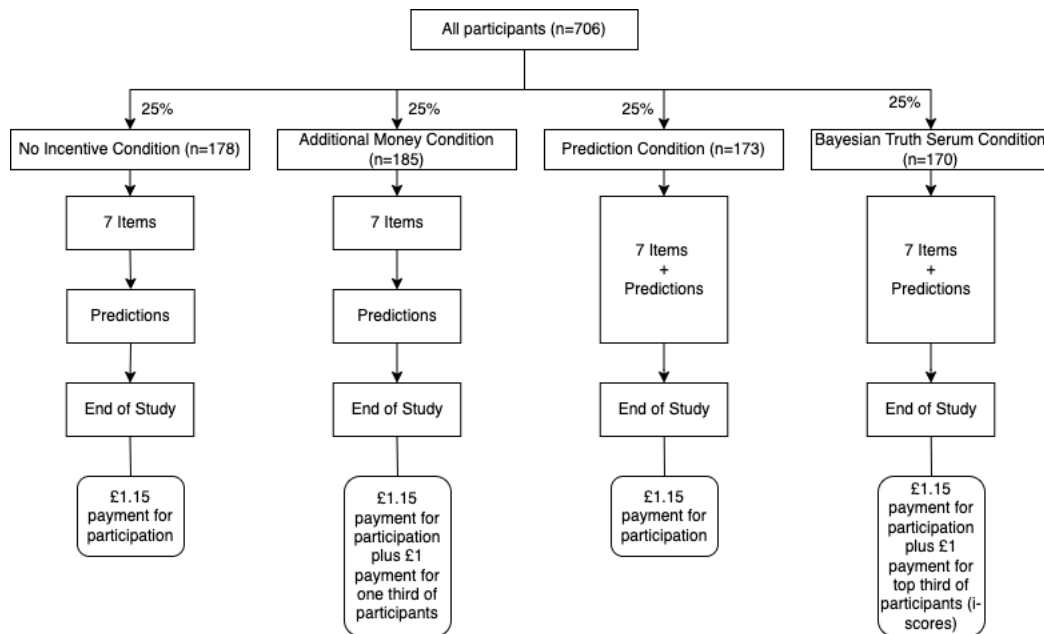

Figure 3. Experimental Outline.

---

[11] 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.

Those in the Bayesian Truth Serum Condition and in the Prediction Condition were asked to also provide predictions as to the underlying distribution of answers on the same page where they provided their own response (see Figure 2 for an illustration). Specifically, they were to provide the frequency of every of the seven answers (1 through 7) to each item, with the constraints that each estimate cannot be smaller than '1' and they all have to sum to '100'. Participants in the No Incentive and Additional Money Conditions were also asked to provide predictions (to hold constant participation payment earnings per time), though they were only to make these predictions once they had provided their own answers to all seven items and had moved on to the next page. Those in the Bayesian Truth Serum Condition and in the Additional Money Condition received £1 additional payment if they were in the top 33% of information scores. Those in the Bayesian Truth Serum condition received 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 were simply told how they would be compensated for their participation, with the exception of those in the Additional Money Condition, who were presented with a formulation similar to the BTS explaining that the top third of quality responses would receive a bonus. The specific wording was: "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".

Results

Pre-Registered Analyses

Our final sample included 706 participants between 18 and 93 years old (*M*=40.37, *SD*=14.86), 55% of which were female. 178 participants were randomly selected into the No Incentive condition (NI), 185 in the Additional Money condition (AM), 173 in the Prediction condition (P), and 170 in the Bayesian Truth Serum condition (BTS). For a graphical overview of their response distributions on all seven items see the boxplot/violin graphs in Figure 4.
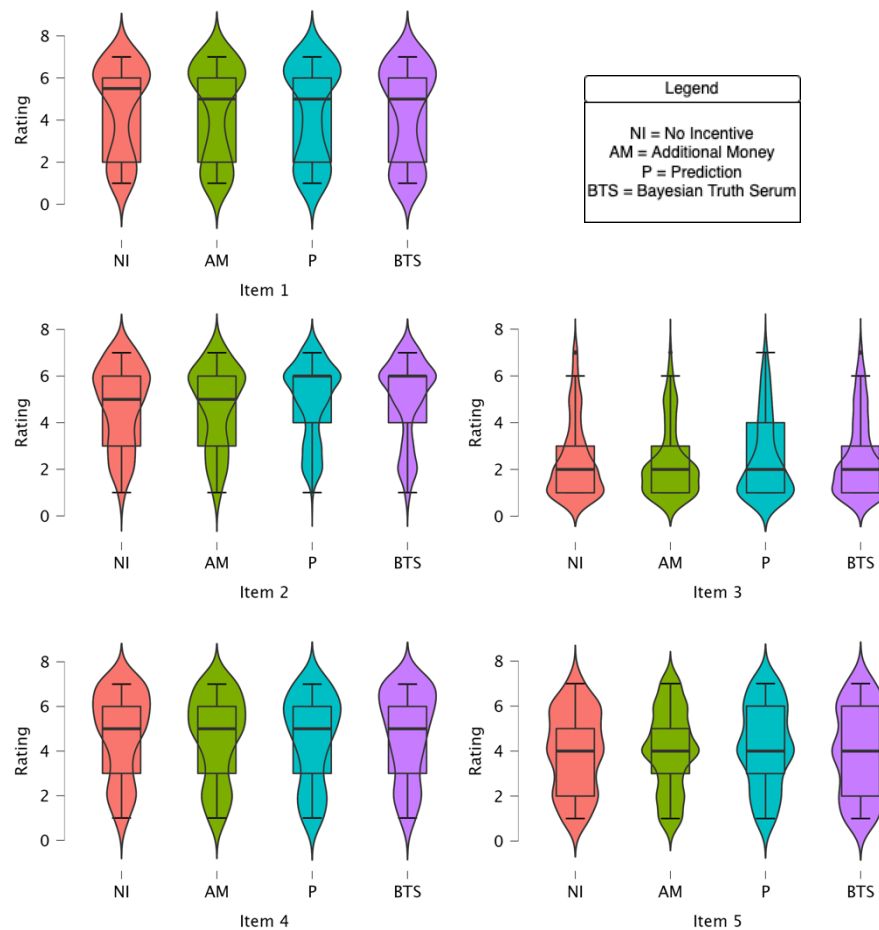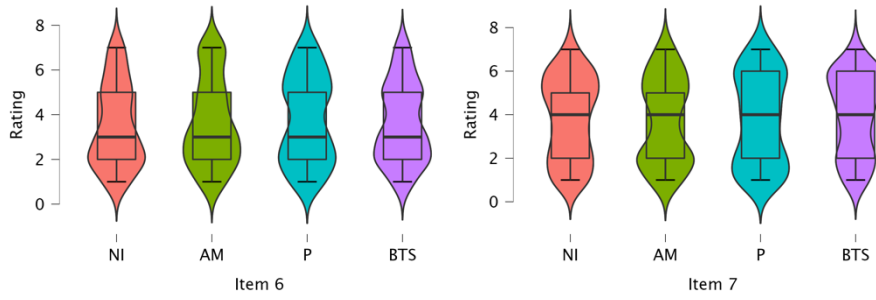
Figure 4. Distribution of Answers by Condition.

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 are interpreted as non-significant. This adjusted significance threshold is used for all tests as all tests include seven comparisons. We treat a pattern of data that shows significant changes in response patterns in at least four out of seven items as strong evidence, as Schoenegger (2021) reported significant differences for four items (at the p<.001 level even though it did not explicitly adjust for multiple comparisons). We treat a pattern that shows significant differences in response patterns in between one and three out of seven items as weak evidence. We treat patterns that show no significant differences as evidence in favour of a null effect of up to V=.1.

We conducted a series of Pearson's χ^2 Goodness-of-Fit tests with the No Incentive condition as the expected distribution and the BTS as the observed distribution, as well as further comparisons that have the BTS as the expected contribution and other treatments as the observed distributions. In our main pre-registered analyses we fail to find significant differences of the Bayesian Truth Serum compared to the No Incentive condition at the .007 level for any of the seven

items (Table 1 [1]).[12] This indicates a replication failure based on our pre-registered criteria. The results from Schoenegger (2021) indicated an average Cramer's V=.117 across all seven items. The mean effect size in this study is V=.087, smaller than in the original paper. We also report pre-registered comparisons between the treatments, where we find a heterogeneous picture. Specifically, we find weak evidence (Items 6 and 7) that the Additional Money condition differs from the BTS (Table 1 [2]), and we also find weak evidence (Item 7) that the Prediction condition differs from the BTS (Table 1 [3]).

TABLE 1—GROUP COMPARISONS

	[1] BTS vs. Control	[2] Additional Money vs. BTS	[3] Prediction vs. BTS
Item 1	6.649 (.081)	4.858 (.066)	12.970 (.112)
Item 2	17.708 (.132)	16.521 (.122)	10.100 (.099)
Item 3	6.077 (.077)	7.331 (.081)	6.035 (.076)
Item 4	5.591 (.074)	8.924 (.090)	9.817 (.097)
Item 5	10.063 (.099)	8.447 (.087)	6.199 (.077)
Item 6	4.547 (.067)	19.792* (.134)	11.571 (.106)
Item 7	6.156 (.078)	20.214* (.135)	23.794* (.151)
Total n	348	355	343

Notes: Pearson's χ^2 Goodness-of-Fit test statistics. Effect size Cramer's V in parentheses. No Incentive condition coded as expected distribution for [1]. BTS condition coded as expected distribution for [2] and [3]. df=6 for [1]-[3].
*p<.007

As pre-registered, we also report χ^2 Tests of Homogeneity as well as Kolmogorov-Smirnov tests as additional analyses in the appendix, where we find no significant differences between any conditions on either test at the .007 level, see Appendix B.

---

[12] Though note that the p-value for Item 2 is .007006, which we interpret as non-significant based on our pre-registered adjusted alpha.

Non-Pre-Registered Analyses

Below, we continue our analyses from above, though all the comparisons below were not pre-registered. First, comparing other treatments to the control, we also do not observe any significant effects in the Additional Money condition (Table 2 [4]), and provide weak evidence for the Prediction condition (Table 2 [5]), where one item (Item 5) shows a significant difference in distributions compared to the No Incentive condition. Importantly though, we also find weak evidence (Items 3 and 6) that the Prediction condition and the Additional Money condition produce significantly different distributions (Table 2 [6]).

TABLE 2—NON-PRE-REGISTERED GROUP COMPARISONS

	[4] Additional Money vs. Control	[5] Prediction vs. Control	[6] Prediction vs. Additional Money
Item 1	5.300 (.069)	15.084 (.121)	8.371 (.090)
Item 2	7.056 (.080)	11.135 (.104)	7.320 (.084)
Item 3	11.617 (.102)	10.578 (.101)	19.909* (.139)
Item 4	3.966 (.060)	10.012 (.098)	2.909 (.053)
Item 5	12.095 (.104)	19.927* (.139)	4.766 (.068)
Item 6	13.623 (.111)	10.016 (.098)	29.766* (.169)
Item 7	15.579 (.119)	15.824 (.124)	15.655 (.113)
Total n	363	351	358

Notes: Pearson's χ^2 Goodness-of-Fit test statistics. Effect size Cramer's V in parentheses. No Incentive condition coded as expected distribution for [4]-[5]. Additional Money condition coded as expected distribution for [6]. df=6 for [4]-[6]. *p<.007

Discussion

The data presented in this paper do not show any significant differences between the Bayesian Truth Serum condition and the No Incentive control condition. As pre-registered, we treat this pattern of data as being evidence in favour of a null effect

of up to Cramer's V=.1 and as such a failure to replicate the results of Schoenegger (2021). However, the current study was not powered to detect all effects of potential interest. Accordingly, we reserve judgment as to whether the BTS mechanism should be adopted in social science fields that rely heavily on Likert-scale items reporting subjective data as we have studied in this context. This reservation of judgement then opens up the space for further research as our inability to recommend the Bayesian Truth Serum as an incentivisation mechanism that ought to be applied widely leaves open the central question of how to properly achieve this task and what the effect of the BTS is in different contexts, for instance, for items that are different in nature than the ones we considered here. It may be that the Bayesian Truth Serum's applicability is more restricted than we anticipated, that another mechanism is better suited for this context, or that the present study was simply not sufficiently powered to detect small but still relevant effects. This is why we argue that, going forward, issues of incentivisation ought to remain central in further (social) scientific reform efforts and we call for more research in this area.

Obvious targets for future research are the type of items or tasks incentivisation mechanisms are likely to affect, the participants who are more or less susceptible to these manipulations, and the ratios of base pay, bonus pay, bonus probability, and study duration that may make for effective interventions (e.g., Bay et al., 2020; Ho et al., 2015). A limitation of our design and our study's generalisability is that our results might be primarily a function of item choice. While we cannot rule this out, we point out that the studied items were quite varied, ranging from moral permissibility judgements to views on causality and determinism. Importantly, previous work (Schoenegger, 2021) had found BTS effects for these specific items. Our results might be at odds with previous studies

that did establish an effect of the BTS (e.g., Frank et al., 2017; Howie et al., 2011; Loughran et al., 2014; Weaver & Prelec, 2013), but we believe they are nevertheless interesting because they show that these effects need not generalize across samples. In this light, it is important to highlight that we recruited participants in the same manner as Schoenegger (2021) did. The main difference between our study and that of Schoenegger's was that we equated the duration of the No Incentive and BTS conditions, by having participants in the No Incentive condition provide predictions after having provided their individual responses, thus fixing the base pay/duration ratio. Participants in Schoenegger (2021) that were assigned to the comparatively shorter No Incentive condition might have felt generously compensated, or conversely, participants in Schoenegger's longer BTS condition might have felt underpaid (but see the next section). Though note that this should not have influenced the results as the additional task was presented after the main variables of interest were collected.

Relating to our secondary hypotheses and in line with previous work (e.g., Buhrmester et al., 2011; Crump et al., 2013; Mason & Watts, 2010; Rouse, 2015) we fail to find a significant effect of the Additional Money condition over the control. We do find a weak effect when comparing the Prediction condition with the control. Supposed knowledge of others' responses (arrived at by invited speculation about response distributions) may thus affect individuals' responses, much like actual knowledge of others' responses does (Lorenz et al., 2011) through a process that is related to de-biasing methods such as counterfactual reasoning (Hoch, 1985; Koriat et al., 1980; Lord et al., 1984) and dialectic bootstrapping (Herzog & Hertwig, 2009; Krueger & Chen, 2014). Although the observation that respondents tend to think that most others would respond like they do (sometimes

referred to as the false consensus effect; Marks & Miller, 1987; Ross et al., 1977) speaks against this thesis; it has been shown that specifically asking about the expectations for one's social circle may make one consider alternative opinions (Galesic et al., 2018). The way participants approach the prediction task is thus likely to influence whether their personal responses are affected by it.

Interestingly, while adding the prediction task to a study may impact response distributions, coupling this with an additional monetary bonus or instructions as found in the BTS may attenuate or ameliorate some of the effects of this constituent prediction part. Some further evidence in favour of this is the finding that the Prediction condition differs from both the BTS and the Additional Money conditions when the treatments are compared directly. As such, while there is no clear picture arising here, we suggest that there may be an effect of combining an increase of expected earnings with the addition of the prediction task that is not equal to the effects of its constituent parts. This poses scientifically interesting and as-of-yet unanswered questions that require further research. For one, these interactions suggest that incentivisation might not be a matter of "one sizes fits all", but that going forward experimenters may need to take into account the diverse experiences, knowledge, and motives participants bring into a study (Camerer & Hogarth, 1999; Sharp et al., 2006).

Conflict of Interest Statement
The authors of this article declare that they have no financial conflict of interest with the content of this article.

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.

Bay, D., Cook, G. L., & Yeboah, D. (2020). Recruiting method and its impact on participant behavior. In K. E. Karim (Ed.), *Advances in Accounting Behavioral Research* (pp. 1-19). Emerald Publishing. https://doi.org/10.1108/S1475-148820200000023001

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.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1), 7-42. https://doi.org/10.1023/A:1007850605129

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.

Galesic, M., Bruine de Bruin, W., Dumas, M., Kapteyn, A., Darling, J. E., & Meijer, E. (2018). Asking about social circles improves election predictions. *Nature Human Behavior, 2*, 187–193. https://doi.org/10.1038/s41562-018-0302-y

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.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237. https://doi.org/10.1111/j.1467-9280.2009.02271.x

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).

Hoch, S.J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 719–731. https://doi.org/10.1037/0278-7393.11.1-4.719

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.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141–155.

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.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Krueger, J. I., & Chen, L. J. (2014). The first cut is the deepest: Effects of social projection and dialectical bootstrapping on judgmental accuracy. *Social Cognition, 32*(4), 315-336. https://doi.org/10.1521/soco.2014.32.4.315

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.

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231–1243. https://doi.org/10.1037//0022-3514.47.6.1231

Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22) 9020-9025. https://doi.org/10.1073/pnas.1008636108

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.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102*(1), 72–90. https://doi.org/10.1037/0033-2909.102.1.72

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.

Ross, L., Greene, D. and House, P. 1977. The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

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

Sharp, E. C., Pelletier, L. G., & Lévesque, C. (2006). The double-edged sword of rewards for participation in psychology experiments. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 38*(3), 269–277. https://doi.org/10.1037/cjbs2006014

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 A

ITEM 1. Charlie needs to learn how to change a lightbulb, and so he goes to the 'how-to' section in his local library. He finds a shelf full of identical looking books titled Home Repair. In each of these books are step-by-step instructions on the way to change a lightbulb—we'll call the way the book describes way 'w'. Unbeknownst to Charlie, all the copies of Home Repair on the shelf are fakes, except for one. Pranksters have placed these copies there, and these fake copies contain mistaken step-by-step instructions on the way to change a lightbulb. Since Charlie does not know this, he reaches up and grabs the copy of Home Repair nearest to him. By sheer luck, he selects the only copy in the entire library that contains genuine and reliable step-by-step instructions for changing a lightbulb, and he reads the correct step-by-step instructions on the way to change a lightbulb. Had Charlie picked up any of the other guides—which he so easily could have—he would have believed the mistaken instructions were correct.

Do you agree with the following statement? 'Charlie knows how to change a lightbulb.'

ITEM 2. Jamie is a very good rollerblader, but she doesn't know that she is very good at rollerblading. When praised for her rollerblading, she says, "Thank you! But I don't think I'm very good at rollerblading."

Do you agree with the following statement? 'When Jamie said, "Thank you! [...]," she was being modest.'

ITEM 3. Many countries have a problem with their citizens not paying taxes, which costs society a considerable amount of money. Some countries have therefore started to send out information to the taxpayers with the encouraging message "To pay your taxes is the right thing to do". The idea with this intervention is to give tax evaders a bad conscience and therefore increase their motivation to pay their taxes.

To what extent do you think that the described policy restricts the individual's freedom of choice?

ITEM 4. A bomb has been planted in a crowded section of a major city and a terrorist is currently in custody. Beyond a reasonable doubt, this terrorist has the information required for successfully defusing the bomb. If the bomb explodes, it is estimated that thousands of people will die.

If torture is successfully employed against the terrorist, he will provide the information needed to defuse the bomb in time. All of the available evidence from previous situations like this indicates that torture has a high probability of making the terrorist provide the needed information to defuse the bomb in time. Alternative investigational and interrogative methods other than torture still remain as options. These methods have a low probability of success.

Given the situation described above, please indicate which rating best describes your judgment of the following statement: 'It is morally permissible to torture the terrorist.'

ITEM 5. Bruno has just finished painting his house. Bruno painted his house the same color as the sky on a clear summer day. Bruno claims his house is blue.

With respect to the case of Bruno, how much do you agree or disagree with the following statement? 'If a claim reports how the world is, then it is true.'

ITEM 6. Mary is the single mother of two: Mark, 7, Sally, 4. Mary works most of the day, and although she is known for being fairly patient and good natured, over the last year she has exhibited some unusually aggressive behavior toward her neighbor. Last week, when she came back from work late at night, she couldn't drive into her garage because her neighbor had blocked her driveway with his new BMW. Enraged, she stepped on the gas pedal and crashed her car into her neighbor's. Unfortunately, her neighbor was still inside the car (it was too dark for anyone to see him), and both his legs were seriously broken in several places. Now he is not only suing her for several thousand dollars, but he's also pressing charges. However, a neurologist examined her brain and discovered that, in the last year, Mary has been developing a rare tumor in her frontal lobe. Since the frontal lobe is necessary for emotional suppression—that is, the capacity to control one's emotions—the neurologist claims that, unlike a healthy person, Mary was completely unable to control her rage and her desire to smash the car. "In fact", he says, "any person with this kind of tumor", facing the exact same situation, would have done exactly what Mary did. She couldn't have done otherwise. "If Mary is found responsible for her actions, she may be sent to a federal medical facility for the next 6 months". There she could receive medical treatment, but she won't be able to see her children. Unfortunately, during that time, they would be living with Social Services, in what might be a much worse environment for them.

Do you agree with the following statement? 'Mary is morally responsible for crashing her car into her neighbour's.'

ITEM 7. Imagine Jim lives in a causally closed universe. In this universe, given the physical state of the universe, the laws of the universe, and the fixity of the past, at any given moment the universe is closed, like a train moving down the tracks. Whenever Jim makes a decision to act in a particular way, it's always the case that he could have acted differently only if something leading up to his decision had been different. In short, at any given moment, there is one and only one choice and action genuinely open to Jim. Moreover, if you knew absolutely everything about both the history of the universe and about Jim, you could always know in advance what Jim is going to decide to do. He is not the only deciding factor when it comes to what he does. Given the way the world was long before Jim was born, everything in his life is in the cards, so to speak. Jim can make choices, but these choices are the only choices open to him. Now, for illustrative purposes, imagine that Jim decides to take his dog for a walk in the park.

Do you agree with the following statement? 'Jim's choices and decisions make a difference in what he does.'

Appendix B

In these additional analyses [7]-[18], we fail to find any significant differences at the .007 level.

APPENDIX TABLE 1—ALL COMPARISONS FOR χ^2 HOMOGENEITY TEST

	[7]	[8]	[9]	[10]	[11]	[12]
Item 1	3.116	2.553	7.877	2.553	7.877	3.998
Item 2	10.263	6.515	5.512	6.515	5.512	5.845
Item 3	2.845	3.852	3.050	3.852	3.050	9.874
Item 4	2.648	4.139	4.934	4.139	4.934	1.445
Item 5	4.450	3.757	3.201	3.757	3.201	2.310
Item 6	2.271	8.464	5.179	8.464	5.179	12.824
Item 7	3.042	9.449	9.712	9.449	9.712	8.338
Total n	348	363	351	355	343	358

Notes: Pearson's χ^2 Test of Homogeneity test statistics. [7]-[9] compare BTS/Additional Money/Prediction to No Incentive respectively. [10] and [11] compare Additional Money and Prediction to BTS respectively, and [12] compares Prediction with Additional Money. df=6 for [7]-[12].

*p<.007

APPENDIX TABLE 2—ALL COMPARISONS FOR KOLMOGOROV-SMIRNOV TEST

	[13]	[14]	[15]	[16]	[17]	[18]
Item 1	.047	.038	.081	.047	.036	.044
Item 2	.112	.031	.071	.088	.041	.046
Item 3	.032	.059	.044	.067	.064	.067
Item 4	.032	.047	.054	.059	.074	.021
Item 5	.041	.055	.067	.055	.065	.054
Item 6	.054	.054	.090	.061	.062	.092
Item 7	.055	.076	.046	.100	.074	.048
Total n	348	363	351	355	343	358

Notes: Independent Samples Kolmogorov-Smirnov D test statistics. [13]-[15] compare BTS/Additional Money/Prediction to No Incentive respectively. [16] and [17] compare Additional Money and Prediction to BTS respectively, and [18] compares Prediction with Additional Money.

*p<.007