Does Truth Pay? Investigating the Effectiveness of the Bayesian Truth Serum with an Interim Payment: A Registered Report

Claire M. Neville[1] *& Matt N. Williams*[1]

Send correspondence to claire.neville.1@uni.massey.ac.nz

Abstract

Self-report data is vital in psychological research, but biases like careless responding and socially desirable responding can compromise its validity. While various methods are employed to mitigate these biases, they have limitations. The Bayesian Truth Serum (BTS; Prelec, 2004) offers a survey scoring method to incentivise truthfulness by leveraging correlations between personal and collective opinions and rewarding 'surprisingly common' responses. This study evaluated the effectiveness of the BTS in mitigating socially desirable responding to sensitive questions and tested whether an interim payment could enhance its efficacy by increasing trust. In a between-subject experimental survey, 877 participants were randomly assigned to one of three conditions: BTS, BTS with Interim Payment (BTS+IP) and Regular Incentive (RI). Contrary to the hypotheses, participants in the BTS conditions displayed *lower* agreement with socially undesirable statements compared to the RI condition. The interim payment did not significantly enhance the BTS's effectiveness. Instead,

[1] School of Psychology, Massey University

response patterns diverged from the mechanism's intended effects, raising concerns about its robustness. As the second registered report to challenge its efficacy, this study casts serious doubt on the BTS as a reliable tool for mitigating SDR and improving the validity of self-report data in psychological research.

Keywords: *Bayesian Truth Serum (BTS), Data integrity, Incentivising truthfulness, Response biases, Self-report data, Sensitive questions, Socially desirable responding (SDR), Survey methodology.*

Introduction

Self-report data is indispensable in psychological research, enabling the exploration of individual differences, attitudes and behaviours (Baldwin, 2000). However, inherent biases such as careless responding and socially desirable responding (SDR) pose significant challenges to the validity of self-report measures (Arthur et al., 2021). Careless responding ranges from inattentiveness to distinct response styles, such as consistently selecting extreme options or agreeing with statements regardless of content (Nichols et al., 1989). SDR involves portraying positive self-descriptions aligned with social norms, influenced by intentional impression management and unconscious self-deception (Paulhus, 1984; 2002). These biases can introduce systematic errors, undermining the construct validity of self-report measures (Flake & Fried, 2020; Lilienfeld & Strother, 2020).

Researchers employ various post hoc methods to mitigate response distortion effects, such as dropping respondents flagged as providing inaccurate answers (e.g., through attention checks) and applying statistical adjustments. However, each approach has its limitations (Arthur et al., 2021; Lee, 2023). Excluding flagged respondents may result in unrepresentative samples and relies on accurately identifying and quantifying the extent of biased responding. This issue extends to implementing statistical adjustments, which risks introducing unintended bias. Thus, it can be argued that rather than mitigating these limitations post-collection, the challenge lies in proactively addressing the intrinsic biases that undermine the reliability of self-report data at the point of collection.

Bayesian Truth Serum

One mechanism that purports to do this is the Bayesian Truth Serum (BTS; Prelec, 2004). The BTS offers a quantitative method for encouraging truthful (responses to subjective questions by scoring the truthfulness of responses and rewarding higher scores with a bonus payment. As the name implies, it draws on Bayesian principles, involving updating beliefs based on new evidence or information. The BTS also capitalises on a well-established cognitive bias wherein individuals tend to overestimate the prevalence of their own views within a population (Choi & Cha, 2019; Mullen et al., 1985; Ross et al., 1977). As a result, others in the population generally underestimate the actual frequency of one's genuine views, such that they are more common than collectively predicted or 'surprisingly common' (for a hypothetical example, see Weaver & Prelec, 2013, pp. 290-291).

The BTS operates by informing participants that the survey uses an algorithm for truth-telling. They are told that the algorithm will assign scores based on the truthfulness of their answers, with the highest ranking scores earning a bonus in addition to the base pay for participation. The specific calculation method is typically not explained. Participants complete the survey, providing personal answers and predicting others' responses to each survey question. At the end of the study, participants receive their base payment, and those with the highest overall scores receive a bonus.

The BTS functions at the level of an individual question, assigning a specific score (BTS score) to each answer. The BTS score combines an information score (i-score) and a prediction accuracy score. Across a study, these scores can be aggregated to provide a total score for each respondent.

The i-score for each answer *k* measures how truthful respondent *r*'s answer is based on how common it is relative to the group's predictions. Answers that are more common than the group collectively predicts (i.e., surprisingly common) receive higher i-scores. The formula for the i-score is:

$$i\ score\ = \sum_{k} x_{kr}\ log\left(\frac{\bar{x}_k}{\bar{y}_k}\right)$$

Where:

- x_{kr} is 1 if respondent *r* chooses answer *k*, and 0 otherwise.

- $\bar{x}_k$ is the actual average frequency of answer *k* given by all respondents.
- $\bar{y}_k$ is the geometric mean of the predicted frequencies for answer *k* made by all respondents.

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

$$Prediction\ Accuracy\ Score\ = \alpha \sum_k \bar{x}_k \log\left(\frac{y_{kr}}{\bar{x}_k}\right)$$

Where:

- α is a constant that fine-tunes the weight given to the prediction error.
- y_{kr} is respondent *r*'s prediction of the distribution for answer *k*.

The BTS score for respondent *r* for answer *k* combines the i-score and the prediction accuracy score to provide an overall score indicating the 'quality' of the response as follows:

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

Several fundamental assumptions underlie the BTS, particularly regarding participants' rational behaviour. Within the framework of the BTS, truth-telling is

considered individually rational, with participants striving to maximise their expected BTS score. This relies on establishing a Bayesian Nash equilibrium, where each participant's strategy is optimised based on their beliefs about others' strategies. In the above equation, a Bayesian Nash equilibrium exists for $\alpha > 0$, and the game is zero-sum for $\alpha = 1$. In this equilibrium, all participants are assumed to tell the truth to maximise their BTS score and earn a bonus, with no incentive to deviate from their chosen strategy unilaterally.

In real-world scenarios, however, individuals may not consistently exhibit the behaviour expected of Bayesian agents (Trautmann & van de Kuilen, 2011), highlighting the importance of validating the BTS through experimental applications. Promisingly, Frank et al.'s (2017) large-scale experiments validated the BTS in scenarios with both known (coin flips, dice rolls) and unknown (pricing survey) honesty distributions. However, applications in economics, marketing, experimental philosophy and psychology have yielded mixed findings. For instance, in experimental philosophy, Schoenegger and Verheyen's (2022) registered report failed to replicate Schoenegger's (2021) findings, where pairwise comparisons revealed significant differences ($p < .001$) in answer distributions between BTS and control conditions. Nonetheless, there is a prevailing notion that the BTS holds promise in fostering more candid responses in various contexts, including those involving sensitive topics (John et al., 2012; Loughran et al., 2014).

In cases where the BTS encounters limitations or lacks support, common explanations point to participants' unfamiliarity with or disbelief in the method (Barrage &

Lee, 2010; Bennett et al., 2018; Menapace & Raffaelli, 2020), reflecting the challenge of engendering trust in a mechanism that operates without explicit explanation. Furthermore, uncertain incentives for truth-telling may compromise the BTS's effectiveness (Bennett et al., 2018), particularly among online respondents who harbour doubts about promises of bonus payments in general. These doubts can lead to the perception of the BTS as little more than cheap talk. Hence, there is a need for experimental applications of the BTS to examine the effects of addressing these potential shortcomings by aiming to enhance trust both in the mechanism itself and in the bonus payment process.

Study Purpose & Hypotheses

This study aimed to evaluate the effectiveness of the BTS in improving the reliability of self-report data in psychology, focusing on mitigating biases associated with sensitive questions. To address potential challenges such as participant scepticism and uncertainty about incentives, we introduced an interim payment midway through the survey. The interim payment was intended to serve a dual purpose: demonstrating the researchers' ability to detect truthful responses and commitment to fulfilling bonus payments. Based on Weaver and Prelec's (2013) findings that participants became more truthful in response to feedback on their earnings, we expected that integrating this payment would make participants perceive both the mechanism and the incentives as more credible, potentially bolstering its efficacy.

In investigating these aims, two BTS experimental conditions were specified: one without an interim payment and one with an interim payment. In both conditions, each participant's BTS score for each item was calculated and summed. As the survey was undertaken in two parts (see 'Procedure' section), the items were summed for each of the two parts of the survey. In the former condition, both bonuses were paid at the survey's conclusion. In the latter condition, bonuses for summed Part 1 scores were paid at the midway point, and bonuses for summed Part 2 scores were paid at the end of the survey, with the midway bonus serving as the interim payment. The Regular Incentive condition served as the control group, where participants received the participation payment without any additional incentives.

The rationale for the study hypotheses was that greater agreement with socially undesirable statements, resulting in higher scores, would indicate more truthful responses. Research supports this expectation, showing that higher prevalence estimates are more valid for assessing sensitive or socially undesirable behaviours (de Jong et al., 2010; Lensvelt-Mulders et al., 2005) and that misreporting undesirable attitudes results from the same distortions as misreporting about behaviours (Tourangeau & Yan, 2007).

Specifically, the study hypotheses were as follows:

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

- H2: Participants subjected to the BTS with an interim payment will have significantly higher mean scores indicating agreement with socially undesirable statements compared with those subjected to the BTS alone.

Method

Design

The study employed a between-subject, experimental survey design. The study design, hypotheses and analysis plan were pre-registered as part of a registered report submission. The approved Stage 1 manuscript is publicly available at [https://osf.io/vuh8b]. Table 1 provides an overview of the study design plan based on the Peer Community In Registered Reports (PCI RR) template (PCI, 2022).

Participants

Participants aged 18 and over from the US, Canada, UK, Ireland, Australia and New Zealand were recruited through Prolific (Prolific, 2024a) to reflect the international scope of this research. This selection ensured linguistic and cultural coherence, enhancing data consistency and comparability. Prescreeners included fluent English proficiency and the completion of at least 20 previous surveys, based on Prolific's data showing that experienced participants are more likely to complete multi-part surveys, thereby reducing attrition (Prolific, 2024b).

Table 1

Study Design Planner

Research Questions	Hypotheses	Sampling Plan	Analysis Plan	Rationale for Test Sensitivity	Interpretation	Theory Relevance
RQ1: Can the BTS effectively incentivise honesty in Likert scale questions prevalent in psychology research?	H1: Participants subjected to the BTS (with or without an interim payment) will have significantly higher mean scores indicating agreement with socially undesirable statements compared with those in the Regular Incentive condition.	A target sample of 876 participants will be recruited through Prolific. This sample size, determined through a power analysis, accounts for a 10% participant exclusion rate based on recent comparable research and considers the 2-part nature of the survey.	To test the hypotheses, planned contrasts (Ψ) will compare mean scores (μ) between groups: Ψ1: BTS (with or without interim payment) vs. Regular Incentive Ψ2: BTS with interim payment vs. BTS alone Bayes factors will be calculated to evaluate potential null effects.	A power analysis suggests that this sample size will have a statistical power of .8 to detect a small effect size of Cohen's f = 0.1 at an adjusted alpha level of .025.	H1 will be considered supported if the mean score is higher in the BTS condition (with or without interim payment) than in the RI condition, with $p < .025$, 1-tailed. H2 will be considered supported if the mean score is higher in the BTS + IP condition than in the BTS condition, with $p < .025$, 1-tailed.	Theoretically, the idea that the BTS (with or without an interim payment) could be used in a psychology research context to elicit truthful responses to self-report questions could be (un)supported by these analyses.
RQ2: Does the inclusion of an interim payment enhance the efficacy of the BTS mechanism?	H2: Participants subjected to the BTS with an interim payment will have significantly higher mean scores indicating agreement with socially undesirable statements compared with those subjected to the BTS alone.					

The predicted effect size, guided by Cohen's conventions (Cohen, 1988), aimed for the smallest meaningful effect, as advised by Lakens (2022). The a priori power analysis targeted a statistical power of .8 to detect a small effect size of Cohen's $f = 0.1$ at an alpha level of .025, accounting for the Bonferroni correction (see 'Primary Analysis' section). This analysis suggested a sample size of 787 participants. To calculate the sample size for a one-sided test with $\alpha = .025$, the 'α err prob' setting was specified at .05 as, by definition, an F-test is undirected. With three conditions, this sample size translated to approximately 263 participants per group. While Schoenegger (2021) estimated a 5% exclusion rate, it was possible that the current two-part study would experience higher attrition. Therefore, with reference to comparable multi-part studies (Kothe & Ling, 2016; Williams et al., 2024), an exclusion rate of 10% was considered more appropriate, leading to an adjusted target sample size of 876 participants (292 per group).

The target of 292 participants for each group was reached shortly after the survey launch. Once this target was met, data collection ceased without a time-based stopping rule. However, as the survey was completed in two parts, a time-based stopping rule was implemented for Part 2 of the survey. Data collection for each group continued until a 72-hour time limit was reached from when the invitation to complete Part 2 was sent.

Procedure

The survey was conducted in two parts. In Part 1, participants were recruited via a short advertisement posted on Prolific. They were then directed to a Qualitrics (Qualitrics, 2024) survey, which began with an information sheet and a consent item. Participants were invited to return approximately 48 hours after Part 1 was closed to complete the second part of the survey. At the conclusion of each part of the survey, participants were automatically directed back to Prolific with a completion code.

Using the randomiser function in Qualitrics, participants were randomly assigned to one of three conditions: 'BTS' (BTS Alone), 'BTS + IP' (BTS with Interim Payment), or 'RI' (Regular Incentive). In all conditions, participants received a total base payment of £1, with £0.50 paid upon completion of Part 1 and £0.50 upon completion of Part 2. These base payments were in line with Prolific's guidelines, converting to an hourly rate of £15 for survey completion. To minimise potential order effects, each main questionnaire item was paired with its associated prediction question, and these pairs were presented in a randomised order to each participant across all conditions.

In the 'BTS' condition, participants first read an adaptation of the BTS text prepared by Frank et al. (2017) before answering questions. This introductory text (Figure 1) clarified that the top 50% of participants, based on their aggregated BTS scores for each part of the survey, would receive a maximum bonus of £1 (£0.50 per part) payable upon survey completion. This bonus amount was based on Schoenegger’s (2021) study. The departure from the conventional 30% allocation in

previous studies aimed to enhance engagement with the survey by offering a greater probability of receiving the bonus while maintaining moderate levels of uncertainty to strengthen motivation. After each question, participants were prompted to predict how others in the study would respond in percentage terms, indicating the expected distribution of responses on the Likert scale. The peer prediction question in Qualitrics dynamically updated to show participants whether their predictions summed up to 100%, streamlining the prediction process and reducing participant effort and time. Participants were ranked within the BTS condition to determine the top 50% eligible for a bonus based on the sum of their BTS scores in each part.

Figure 1

'BTS' Condition Instructions

Work by MIT researchers 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 a score to your responses which indicates how truthful and informative you are being. Once we have collected all of the responses to Part 1 of this survey, we will rank the survey responders by the sum of their truthfulness scores and award a bonus of £0.50 to the responders in the top 50%. The process will repeat for Part 2, following a separate invitation from Prolific to complete the survey. You will be notified of whether you have earned a bonus only after Part 2 has been completed. These bonuses, along with your base pay for participation, will be paid at the end of the study.

In the 'BTS + IP' condition, participants followed a process similar to that of the 'BTS' condition. They started by reading an adaptation of the BTS text (Figure 2) specific to their condition, which explained that the top 50% of participants, based on the sum of their BTS scores for Part 1, would receive a partial bonus of £0.50, payable after Part 1. Similarly, the top 50% in the condition, based on the sum of their BTS

scores for Part 2, would receive a partial bonus of £0.50, payable after Part 2. The bonus payment after Part 1 constituted the ‘interim payment’. After answering each question, participants made peer predictions. Participants were ranked within the BTS + IP condition to determine the top 50% eligible for a bonus in each part.

Participants in the ‘RI’ condition did not receive a BTS text. However, to maintain consistent base compensation per hour across conditions, they made predictions following the main questions as in the two BTS conditions.

Figure 2

‘BTS + IP’ Condition Instructions

Work by MIT researchers 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 a score to your responses, which indicates how truthful and informative you are being. Once we have collected all of the responses to Part 1 of this survey, we will rank the survey responders by the sum of their truthfulness scores and pay a bonus of £0.50 to the responders in the top 50%. The process will repeat for Part 2, following a separate invitation from Prolific to complete the survey. These bonus payments are in addition to the base pay for participation.

Attention Checks

Attention checks were not included in the study, as it aimed to assess the effectiveness of the BTS in addressing both socially desirable and careless responses. This ensured that any improvements in accuracy were solely attributed to the BTS rather than the presence of attention checks.

Quality Check

In Part 2 of the survey, participants in each of the BTS conditions were asked, "What percentage of participants, ranked by their BTS scores, will be eligible for a bonus?". The options were 30%, 50%, or 100%. This manipulation check ensured that participants understood that the top 50% in their respective group would be eligible for a bonus. Participants were only provided information about the specific condition to which they were assigned.

Measures

The main questionnaire comprised ten questions addressing potentially uncomfortable or sensitive topics. These questions were curated from established scales, developed or updated in the 21st century, including the Social Dominance Orientation (SDO7; Ho et al., 2015), Belief in Sexism Shift (BISS; Zehnter et al., 2021), and Succession, Identity, and Consumption Scale of Prescriptive Ageism (SIC; North & Fiske, 2013). Table 2 lists the survey questions and source scales. The use of contemporary scales ensured relevance to current social attitudes and sensitivities. The study questionnaire employed a consistent 5-point Likert scale format for all items.

Table 2

Survey Questions and Sources

Item Number	Item Statement	Source
Q1	Some groups of people are simply inferior to other groups.	SD07
Q2	An ideal society requires some groups to be on top and others to be on the bottom.	SD07
Q3	Group equality should not be our primary goal.	SD07
Q4	It is unjust to try to make groups equal.	SD07
Q5	All in all, men have more responsibilities and fewer benefits.	BISS
Q6	Nowadays, men don't have the same chances in the job market as women.	BISS
Q7	Men are not particularly discriminated against. (R)	BISS
Q8	Doctors spend too much time treating sickly older people.	SIC
Q9	Younger people are usually more productive than older people at their jobs.	SIC
Q10	Older people don't really need to get the best seats on buses and trains.	SIC

Note. SDO7 = Social Dominance Orientation; BISS = Belief in Sexism Shift; SIC = Succession, Identity, and Consumption Scale of Prescriptive Ageism. (R) indicates reverse-coded item.

By selecting questions from a range of constructs, the BTS was tested for its ability to elicit truthful responses across various dimensions in aggregate. In each condition, responses to all ten questions were combined into a single social undesirability score for each participant. Cronbach's alpha, calculated across imputed datasets, showed moderate reliability with a mean of .618 (*SD* = 0.004). The choice of ten main questions sought to balance thorough data collection with the need to keep the survey manageable and engaging for participants, taking into account the additional

onus of prediction tasks. This approach sought to ensure fair compensation and avoid participant fatigue, aligning with budget constraints and guidelines for survey length (Denison, 2023).

The survey also included various demographic items, including age bracket, gender and education level. The survey questionnaire can be viewed here.

Ethics

This study was approved by the Massey University Human Ethics Committee (MUHEC).

Analysis Strategy

The analyses were conducted using R (R Core Team, 2024) after data cleaning. Missing data was handled by performing multiple imputations using the `mice` package in R (van Buuren & Groothuis-Oudshoorn, 2011), following Rubin's (1987) guidelines. Five imputed datasets were generated with a proportional odds model for ordered categorical variables. Statistical analyses were performed on each imputed dataset separately, and results were combined using Rubin's Rules via the pool() function in `mice` (van Buuren, 2018).

Descriptive Analysis

Descriptive statistics were provided to summarise the sample characteristics in terms of age group, gender and education. These data were not used in hypothesis testing but served solely to describe the sample.

Primary Analysis

Planned contrasts (Ψ) were used to test the hypotheses, allowing for specific, theory-driven comparisons between groups based on prior expectations (Field, 2018). While the preregistration specified a Welch adjustment to address variance inequalities (Zimmerman, 2010), the use of linear models with planned contrasts instead of t-tests per se, combined with the need to pool variance estimates across the five imputed datasets, rendered this approach impractical. Instead, HC3 robust standard errors, a heteroscedasticity-consistent covariance matrix (HCCM), were applied as a suitable alternative (Long & Ervin, 2012). To present the two closest alternatives to the original method, Welch-adjusted t-tests were also conducted separately on each imputed dataset, with detailed results reported in the Supplementary Materials (Table S0).

The contrasts compared:

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

Weights were assigned as follows:

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

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

Orthogonality was confirmed by the sum of the products of the weights equaling zero, ensuring each contrast tested a distinct hypothesis. To control the familywise Type I error rate, we applied a Bonferroni correction (Bonferroni, 1936) by dividing the alpha level by the number of contrasts. Thus, the alpha level was set at $\alpha = 0.025$ for each test. While Cohen's *f* informed the a priori power analysis, Cohen's *d* was calculated during the analysis to quantify effect sizes for the pairwise planned contrasts.

The following inferential criteria applied:

- H1 will be considered supported if the mean score is higher in the BTS condition (with or without interim payment) than in the RI condition, with $p < .025$, 1-tailed.
- H2 will be considered supported if the mean score is higher in the BTS + IP condition than in the BTS condition, with $p < .025$, 1-tailed.

Supplementary Analysis

Bayes factors were calculated using the `BayesFactor` package in R (Morey et al., 2018) to compare non-directional alternatives of the original hypotheses against zero-effect null hypotheses through direct group comparisons. The default Cauchy prior (scale parameter 0.707) was used for the effect size under the alternative hypothesis. Calculations were averaged across imputed datasets (Hoijtink et al., 2019a). Bayes factors were interpreted contextually, with values around 1 suggesting no preference between hypotheses. Following guidance from Hoijtink et al. (2019b, p. 545, *How Large*

Should the Bayes Factor Be?), we considered Bayes factors as direct and quantitative indicators of the evidence for (or against) the alternative hypothesis in comparison to the null hypothesis rather than applying strict thresholds. ~~Bayes factors indicated the strength of evidence for each hypothesis, with values around 1 suggesting no preference between hypotheses and other values interpreted contextually. We avoided fixed thresholds, following Hoijtink et al.'s (2019b, p. 545) guidance to view Bayes factors as relative indicators rather than strict criteria.~~ While this supplementary analysis did not influence the determination of the main hypotheses, it provided additional context to determine whether non-significant results in the primary analysis are more consistent with a true null effect or a potential backfire effect.

Exploratory Analysis

To gain further insights, Chi-square tests of independence were undertaken to examine the distributions of individual item responses, cross-tabulated with condition. Post hoc analyses, including Brown-Mood median tests and Welch's t-tests for response durations, were also performed to better understand item-level variability and unexpected effects. These analyses are reported in the supplementary materials.

Outcome Neutral Tests

As preregistered, findings would be considered inconclusive if more than 50% of participants failed to identify the bonus allocation percentage during the manipulation check in the 'BTS' and 'BTS + IP' conditions. 95.94% of BTS participants and 76.00% of

BTS+IP participants correctly identified the allocation, surpassing the 50% threshold for conclusive results.

Results

In total, 877 participants were included in the study and assigned to one of three conditions: BTS (*n* = 289), BTS+IP (*n* = 293) and RI (*n* = 295). The sample's age distribution spanned a broad range, with 68% falling between 25 and 44 years of age. The median age group was 25–34 years. Gender distribution included 59% identifying as female, 39% as male and 1% as non-binary. 40% of participants held a bachelor's degree, and 19% reported a graduate or professional degree, indicating a strong representation of higher education in the sample. Pooled means, 95% confidence intervals (CIs), and standard deviations of participants' social undesirability scores are presented in Table 3.

Table 3

Descriptive Statistics of Social Undesirability Scores by Condition (Post-Imputation)

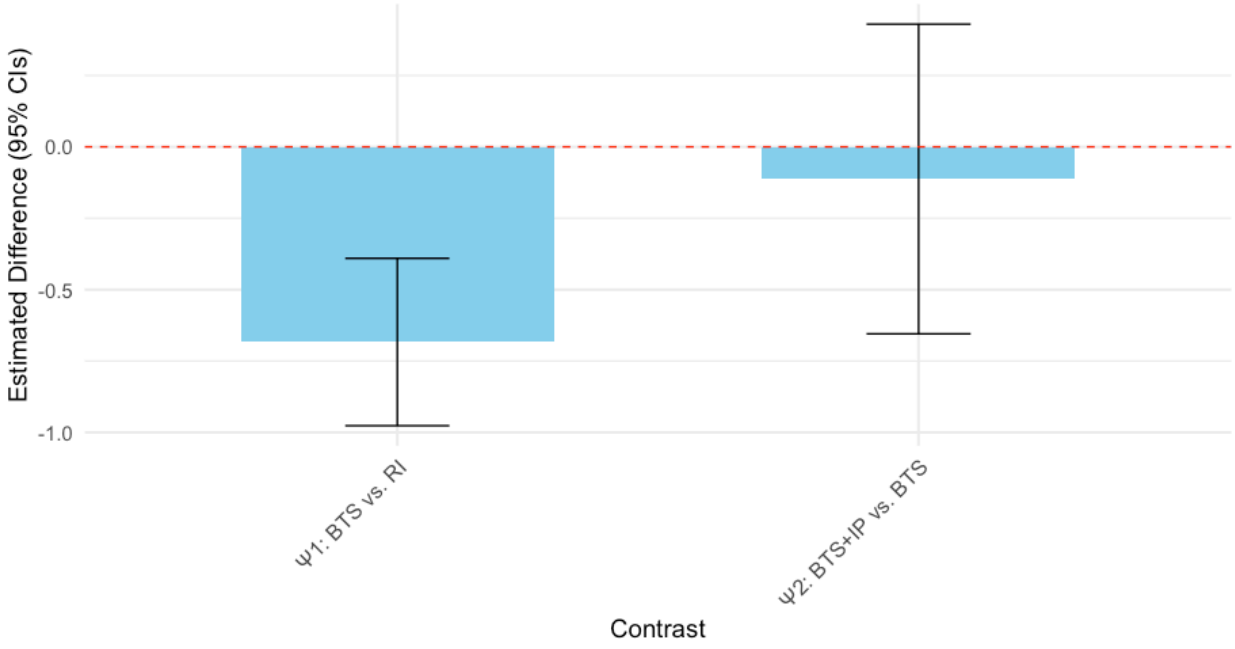
Condition	Mean	95% CI Lower	95% CI Upper	Standard Deviation
BTS	23.1	22.4	23.7	5.71
BTS+IP	22.9	22.2	23.5	5.83
RI	25.0	24.4	25.6	5.28

Note. BTS = Bayesian Truth Serum; RI = Regular Incentive; IP = Interim Payment. Pooled means, confidence intervals (CIs) and standard deviations are pooled across five imputed datasets. CIs are unadjusted 95% intervals for descriptive purposes.

Primary Analysis: The first planned contrast compared the BTS conditions (with or without an interim payment) to the RI condition. It did not support the hypothesised directional effect, $t(848.02) = -5.23$, $p = 1.00$ (one-tailed), $d = -0.36$ (95% CI [−0.49, −0.22]). The mean difference ($M = -0.68$, (95% CI [−0.98, −0.39]) indicates that agreement with socially undesirable statements was lower in the combined BTS conditions compared to the RI condition, contrary to Hypothesis 1. As a one-tailed test was preregistered, this result is interpreted within that framework. However~~Notably~~, the result would have been significant if a two-tailed test had been pre-registered. The second planned contrast compared the BTS+IP condition to the BTS alone. It was also not significant, $t(792.70) = -0.47$, $p = .68$, (one-tailed), $d = -0.03$ (95% CI [−0.17, 0.11]), thereby failing to support Hypothesis 2. These findings are depicted in Figure 3. A supplementary analysis using Welch-adjusted t-tests on each imputed dataset yielded consistent results (see supplementary materials).

Figure 3

Mean Differences and 95% Confidence Intervals for Planned Contrasts



Note. Ψ1 represents the planned contrast comparing BTS (with or without interim payment) vs. Regular Incentive. Ψ2 represents the planned contrast comparing BTS with interim payment vs. BTS alone. Bars represent the estimated difference for each planned contrast, with error bars indicating 95% confidence intervals. The dashed red line represents the null value (0), indicating no difference between conditions.

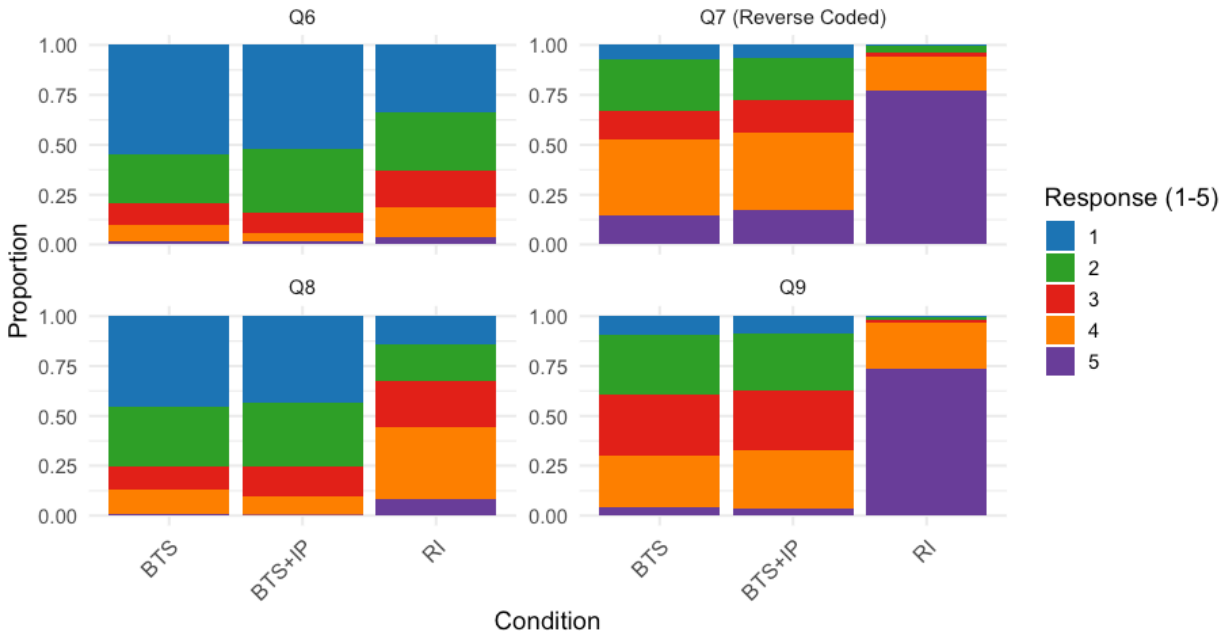
Supplementary Analysis: The Bayesian analysis used Bayes factors (BFs) to compare non-directional alternatives (H1) against zero-effect null hypotheses (H0). For the first contrast, the pooled BF_{10} was 24,757, indicating substantial evidence for a non-null effect (albeit in the opposite direction to that expected). For the second contrast, the pooled BF_{10} was 0.103, suggesting greater consistency with the null hypothesis. These findings align with the primary analysis. Given the supplementary status of this analysis and the lack of a strong basis for prior probabilities, we did not

convert the Bayes factors to posterior probabilities, but interested readers could do so by multiplying the Bayes factors by their own choice of prior odds.

Exploratory Analysis – Chi-square Tests: The preregistered exploratory analysis showed significant associations between condition and response distribution for four of the ten survey items after applying a Bonferroni-corrected significance threshold (α = .005). For three items (Q6, Q8, Q9; see Table 2 for item descriptions), response distributions in the BTS conditions skewed toward positions associated with greater social desirability compared to the RI condition. In contrast, the response distribution for Q7 aligned with the intended effect of the BTS mechanism. These patterns are visualised in Figure 4. No significant associations were observed for Q1–Q5 or Q10. This analysis used unimputed data, as imputing categorical variables can distort frequency distributions (Allison, 2001; van Buuren, 2018). Missing responses (NA) were retained but excluded from the Chi-square calculations.

Exploratory Analysis – Median Tests: Post hoc analysis using the Brown-Mood test (Brown & Mood, 1951) identified significant median differences across conditions for Q6 (χ^2= 38.75, p < .001), Q7 (χ^2 = 296.92, p < .001), Q8 (χ^2 = 142.06, p < .001) and Q9 (χ^2 = 452.43, p < .001). Descriptive analyses of observed medians are presented in Figure 5.

Figure 4

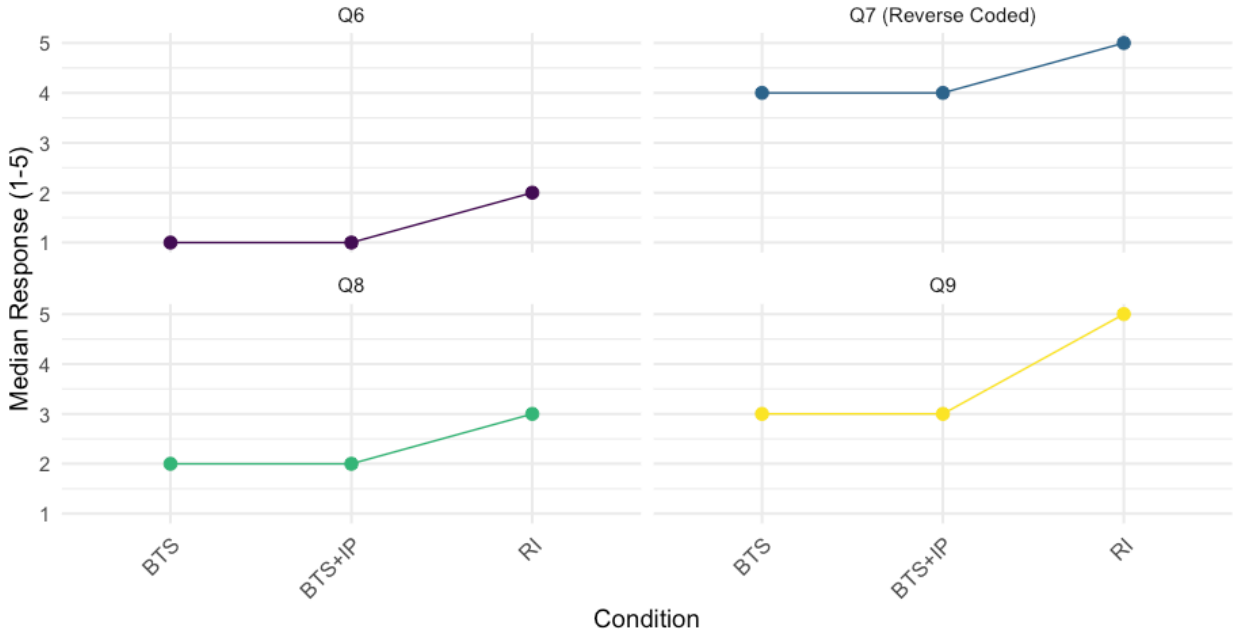


Response Distributions for Items with Significant Differences Across Conditions

Note. Proportions of responses for each condition (BTS, BTS+IP and RI) are displayed for survey questions with significant associations between condition and response distribution. See Table 2 for descriptions of the survey items. Likert scale responses range from 1 (strongly disagree) to 5 (strongly agree), with colour coding indicating response levels.

Figure 5

Median Responses for Items with Significant Differences Across Conditions



Note. Median responses are displayed by condition (BTS, BTS+IP and RI) for survey questions showing significant differences in central tendency. See Table 2 for descriptions of the survey items. Likert scale responses range from 1 (strongly disagree) to 5 (strongly agree).

Exploratory Analysis—Response Durations: Longer total survey response times were examined as a proxy for increased cognitive engagement in the BTS conditions. For Part 1, BTS participants spent significantly more time completing the survey than RI participants ($t(536.43) = 2.67$, $p = .008$), while no significant differences were found between BTS+IP and RI participants ($t(571.42) = 0.90$, $p = .37$). For Part 2 and total duration, no significant differences were observed between conditions.

Discussion

This study evaluated the effectiveness of the BTS in reducing biases in self-reported responses to sensitive questions within a psychological context. The primary analyses, based on pre-registered directional hypotheses, did not support the predicted positive effects of the BTS mechanism in either contrast. Specifically, the planned contrasts failed to reach statistical significance at the pre-registered alpha level, providing no evidence for an increase in agreement with socially undesirable statements in the BTS conditions.

The first hypothesis predicted that participants in the BTS conditions (with or without an interim payment) would exhibit higher agreement with socially undesirable statements than those in the RI condition, thereby reflecting greater truthfulness. However, the first planned contrast revealed no significant effects in the hypothesised direction. Instead, findings indicated that participants in the BTS conditions reported lower agreement with socially undesirable statements than those in the RI condition. Supplementary Bayesian analyses tested non-directional hypotheses against a zero-effect null, revealing substantial evidence for a non-null effect, albeit in the opposite direction to the preregistered predictions. This pattern may indicate a possible backfire effect, wherein the BTS appeared to increase social desirability bias.

Two main explanations are considered to account for this finding. First, the BTS mechanism may have broken down, ~~with the mechanism~~ causing participants to prioritise SDR over truthfulness. This could reflect a failure of the foundational

assumption that participants act as rational agents. Instead, participants may have strategically adopted SDR as their optimal strategy, possibly influenced by factors such as experimenter demand effects or insufficient incentives. For example, participants may have aligned their responses with perceived researcher expectations, knowing their answers would be scrutinised as part of bonus allocation. Similarly, while the bonus amounts used in this study were consistent with those shown to be effective in Schoenegger's (2021) study, they may have been inadequate in this context to offset the perceived costs of truthfulness, such as time, cognitive effort or discomfort associated with disclosing sensitive information (Smith et al., 2014).

Second, the relationship between increased truthfulness and SDR may be more complex than initially assumed, with truthful responses not always reducing SDR. In some cases, truthful responses may align with socially desirable positions rather than contradict them. For instance, agreement with the statement "Younger people are usually more productive than older people at their jobs" may reflect a widely accepted societal norm within a relatively young and highly educated sample rather than a socially undesirable position, as initially assumed. In such cases, lower agreement in the BTS conditions could indicate deeper engagement and a willingness to challenge reflexive, norm-aligned responses. Nesting within this broader complexity, we, the researchers, may have misjudged the direction of SDR for certain items. While these interpretations offer plausible explanations for the observed response patterns, they remain tentative, particularly given the absence of consistent evidence for increased cognitive engagement in the BTS conditions as measured by survey completion times.

The second hypothesis posited that an interim payment would enhance the BTS mechanism's efficacy by increasing participants' trust in the bonus allocation process and the perceived credibility of the incentives. This prediction was also not supported, with no significant difference observed between the BTS and BTS+IP conditions in either the primary or supplementary analyses.

Several factors may explain this result. For instance, the 48-hour timeframe for processing interim bonus payments may have reduced their intended effect. Psychological theories of reinforcement emphasise the power of immediate rewards (Skinner, 1953). While the delay was necessary to ensure the completion of Part 1 and accurate bonus allocation under the BTS mechanism, it may have reduced the salience of the payment and its ability to reinforce trust in the process (Singer & Ye, 2013). Furthermore, confusion about the bonus allocation process may have undermined the interim payment's efficacy, evidenced by the 20% lower manipulation check success rate in the BTS+IP condition compared to the BTS condition. Participants may, for example, have perceived the interim payment as a standalone bonus for completing Part 1 rather than as reinforcement of the broader BTS incentive structure, limiting its intended impact. Alternatively, the BTS mechanism's efficacy may be inherently unaffected by interim payments. Participants may have already trusted the researchers' ability and commitment to pay bonuses without requiring a demonstration thereof, challenging prior assumptions that the mechanism's limitations arise from issues of trust and credibility (Barrage & Lee, 2010; Bennett et al., 2018; Menapace & Raffaelli, 2020).

If participant trust was already established, the interim payment might not have provided any additional benefit.

This study's findings cast serious doubt on the effectiveness of the BTS in improving the accuracy of self-report data, particularly in reducing response biases to sensitive questions. While prior studies have reported promising results in experimental contexts (e.g., John et al., 2012; Weaver & Prelec, 2013), this study, alongside the earlier registered report by Schoenegger and Verheyen (2022), found no evidence for the hypothesised benefits of the BTS. Instead, patterns inconsistent with the mechanism's intended effects, including possible backfire effects, emerged, raising concerns about its robustness. Although further research may uncover specific conditions or refinements that improve its performance, the current evidence does not support the efficacy of the BTS in enhancing truthfulness in applied psychological research.

This study acknowledges several limitations that suggest potential directions for future research. First, the convenience sample, predominantly aged 25–44 and highly educated, limits the generalisability of the findings. Future studies should prioritise recruiting more diverse and representative samples to evaluate the BTS across varied populations and contexts. Furthermore, this study made assumptions about what constitutes a socially desirable stance. However, these assumptions regarding the direction of SDR may not have accurately aligned with participants' norms or beliefs for certain items. Future research could explicitly test these assumptions to ensure they are contextually appropriate and reflect the studied population. While the primary

confirmatory analysis did not aim to test backfire effects, the supplementary Bayesian analysis identified an unexpected pattern that could indicate increased social desirability bias under the BTS. Additionally, the exploratory analyses of item-level distributions provided useful insights, though their post hoc nature limits the strength of the conclusions. These exploratory findings suggest that future preregistered studies should not only address how well social desirability assumptions align with participant norms but also investigate potential backfire effects using a targeted confirmatory approach.

~~Finally, while exploratory analyses yielded useful insights, their post hoc nature limits the strength of the conclusions. Future preregistered studies should explicitly test hypotheses about SDR disruption and norm alignment to better understand the contexts in which the BTS is most effective.~~

Conclusion

This study evaluated the effectiveness of the BTS in reducing response biases and improving the reliability of self-report data in psychological research. Contrary to predictions, participants in the BTS conditions reported lower agreement with socially undesirable statements compared to those in the RI condition, raising concerns about its intended effects. Additionally, the interim payment, designed to enhance trust in the BTS, failed to produce any meaningful improvement. As the second registered report that has found no robust support for the BTS, these findings cast further doubt on its efficacy as a mechanism for eliciting truthful responses in self-report studies. Until

further research identifies conditions under which the BTS performs effectively, it cannot be recommended as a practical tool for applied psychological research.

Acknowledgments

This study was supported by the Marsden Fund Council from New Zealand Government funding, managed by Royal Society Te Apārangi. The Massey University School of Psychology Postgraduate Research Fund provided additional support. We acknowledge using Qualtrics (provided through Massey University) for survey administration and Prolific for participant recruitment. Data analysis was conducted using R (R Core Team, 2024). We also used Grammarly for language proofreading and OpenAI's ChatGPT to refine and troubleshoot software code used in data analysis. This assistance did not influence the scientific interpretation of the results. We thank the reviewers at Peer Community In for their constructive feedback.

Conflict of Interest Disclosure

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

References

Allison, P. D. (2002). *Missing data*. Sage Publications.

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Baldwin, W. (2000). Information no one else knows: The value of self-report. In A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman, & V. Cain (Eds.), *The Science of self-report: Implications for Research and Practice* (pp. 3–7). Lawrence Erlbaum Associates.

Barrage, L., & Lee, M. S. (2010). A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics Letters*, *106*(2), 140–142. https://doi.org/10.1016/j.econlet.2009.11.006

Bennett, R., Balcombe, K., Jones, P., & Butterworth, A. (2018). The benefits of farm animal welfare legislation: The case of the EU broiler directive and truthful reporting. *Journal of Agricultural Economics*, *70*(1), 135–152. https://doi.org/10.1111/1477-9552.12278

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

Brown, G. W., & Mood, A. M. (1951). On median tests for linear hypotheses. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 159–166. https://doi.org/10.1525/9780520411586-013

Choi, I., & Cha, O. (2019). Cross-cultural examination of the false consensus effect. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02747

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

De Jong, M. G., Pieters, R., & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*(1), 14–27. https://doi.org/10.1509/jmkr.47.1.14

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

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

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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

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

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

Hoijtink, H., Gu, X., Mulder, J., & Yves Rosseel. (2019a). Computing Bayes factors from data with missing values. *Psychological Methods*, *24*(2), 253–268. https://doi.org/10.1037/met0000187

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019b). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

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. https://doi.org/10.1177/0956797611430953

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

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.33267

Lee, J. J. (2023). Cheap Talk with the Bayesian truth serum. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4450528

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-Analysis of randomized response research. *Sociological Methods & Research*, *33*(3), 319–348. https://doi.org/10.1177/0049124104268664

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4). https://doi.org/10.1037/cap0000236

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

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. https://doi.org/10.1007/s10940-014-9219-4

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

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *BayesFactor: Computation of Bayes factors for common designs*. R-Packages. https://cran.r-project.org/package=BayesFactor

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, *21*(3), 262–283. https://doi.org/10.1016/0022-1031(85)90020-4

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239–250.

https://doi.org/10.1002/1097-4679(198903)45:2%3C239::aid-jclp2270450210%3E3.0.co;2-1

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

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp. 49–69). Lawrence Erlbaum Associates.

Peer Community In [PCI]. (2022). *Guide for authors*. Rr.peercommunityin.org. https://rr.peercommunityin.org/help/guide_for_authors#h_2751396573533161330 9625021

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

Prolific. (2024a). https://www.prolific.com

Prolific. (2024b). *How do I set up a longitudinal / multi-part study?* https://researcher-help.prolific.com/hc/en-gb/articles/360009222733-How-do-I-set-up-a-longitudinal-multi-part-study#h_01HD485SB6AFZZWTJRM37EYTCR

Qualtrics. (2024). Qualtrics. https://www.qualtrics.com

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-x

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

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

Schoenegger, P., & Verheyen, S. (2022). Taking a closer look at the Bayesian truth serum. *Experimental Psychology*, *69*(4), 226–239. https://doi.org/10.1027/1618-3169/a000558

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112–141. https://doi.org/10.1177/0002716212458082

Skinner, B. (1953). *Science and human behavior.* Macmillan.

Smith, E., Mackie, D., & Claypool, H. (2014). *Social psychology* (4th ed.). Taylor & Francis Group.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Trautmann, S. T., & van de Kuilen, G. (2014). Belief elicitation: A horse race among truth serums. *The Economic Journal*, *125*(589), 2116–2135. https://doi.org/10.1111/ecoj.12160

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press, Taylor & Francis Group.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

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

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

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

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