Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Does Truth Pay? Investigating the Effectiveness of the Bayesian Truth Serum with an Interim Payment

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

in

Psychology

At Massey University, Albany,

New Zealand.

Claire Marie Neville

February 2025

Abstract

Self-report data is vital in psychological research, but biases like careless and socially desirable responding (SDR) can compromise its validity. SDR is particularly problematic for sensitive questions, which are common in psychological studies and prone to distortion (Krumpal, 2013). While various methods have been developed to mitigate response biases, they each have limitations. Many proactive approaches rely on intrinsic motivation, which may be insufficient when questions are sensitive, cognitively demanding or socially undesirable (Crowne & Marlowe, 1964). The Bayesian Truth Serum (BTS; Prelec, 2004) offers an incentive-compatible alternative, aligning extrinsic motivation with truthfulness. It encourages considered and honest responses to subjective questions by scoring the truthfulness of responses and rewarding higher scores with a bonus payment. This approach leverages a Bayesian interpretation of population distributions and ties incentives to how 'surprisingly common' participants' responses are. However, prior research has yielded mixed results, highlighting challenges such as participant scepticism and uncertainty about incentives. This study evaluated the effectiveness of the BTS in mitigating SDR in sensitive questions. It tested whether an interim payment could enhance the efficacy of the BTS 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, 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. Beyond its empirical contributions, this thesis also reflects on the challenges and benefits of adopting the registered report format. Situated within meta-psychology, the reflection explores key moments of academic and personal growth. It may serve as a resource for students and supervisors considering the registered report approach in postgraduate research.

Acknowledgements

This research was supported by the Marsden Fund Council, which received funding from the New Zealand Government and is managed by the Royal Society Te Apārangi. The Massey University School of Psychology Postgraduate Research Fund provided additional support. I also wish to acknowledge the recommender and reviewers at Peer Community In for their insightful and constructive feedback on the Stage 1 manuscript.

I am deeply grateful to my supervisor, Dr Matt Williams, for his unwavering support, guidance and mentorship throughout this research project and for encouraging me to follow the registered report approach. I also want to give special thanks to Dr Pita King for his assistance with the cultural aspects of the ethics application.

Finally, I would like to thank my family and friends, especially Jimmy, whose unwavering support and patience were crucial in completing this work.

Table of Contents

List of Tables and Figures

Introduction

Background

Self-report data collected via questionnaires and surveys is indispensable in psychology research, offering a direct window into individuals' internal states and personal experiences (Baldwin, 2000). These methods are often the only viable option for studying psychological constructs that are inherently subjective and not easily accessible through alternative means. However, reliance on self-report data brings significant challenges, particularly its susceptibility to response biases like careless and socially desirable responding (SDR).

Careless responding encompasses behaviours ranging from inattentiveness to distinct response styles, like consistently selecting extreme options or agreeing with statements regardless of content (Nichols et al., 1989). SDR occurs when individuals intentionally or unconsciously present themselves in a more socially acceptable light (Paulhus, 1984; 2002). Sensitive questions are particularly prone to SDR (Krumpal, 2013). These biases can systematically distort data, reducing the validity of self-report measures and compromising research outcomes (Arthur et al., 2021; Flake & Fried, 2020; Lilienfeld & Strother, 2020; Paulhus & Vazire, 2007).

Addressing response biases is crucial for maintaining the rigour of psychology research. Non-incentive-based methods, grounded in psychological principles, have shown promise in promoting truthful responses in certain contexts (e.g., Fisher & Tellis, 1998; Jacquemet et al., 2013; Rasinski et al., 2005). These approaches primarily rely on intrinsic motivations, such as enjoyment, interest or personal fulfilment (Deci & Ryan, 1985). However, intrinsic motivation alone may be insufficient, particularly when

questions are sensitive, cognitively demanding or socially undesirable (Crowne & Marlowe, 1964). In such cases, extrinsic motivation, involving external incentives like financial rewards, provides a complementary approach by explicitly aligning participants' self-interest with research goals (Prelec, 2004). Techniques such as incentive-based methods seek to address the cognitive, temporal and emotional costs associated with truthfulness, offering a proactive solution to bias mitigation (Schoenegger, 2021; Weaver & Prelec, 2013).

One promising incentive-based method is the Bayesian Truth Serum (BTS; Prelec, 2004), which encourages truthful responses—in the sense of being both carefully considered and honest—to subjective questions by scoring truthfulness and rewarding higher scores with bonus payments. This mechanism leverages Bayesian principles and a cognitive bias wherein individuals overestimate the prevalence of their own views in a population (Choi & Cha, 2019; Mullen et al., 1985; Ross et al., 1977). While theoretically sound, the BTS assumes that participants will act rationally, evaluating incentives and selecting the option that maximises their expected utility. However, real-world decision-making often deviates from these assumptions, with cognitive biases, heuristics and emotional influences potentially disrupting the alignment between incentives and truthful reporting (Trautmann & van de Kuilen, 2011). Consequently, experimental validation is crucial to assess its practical effectiveness.

Applications of the BTS in psychological contexts have shown promising results (John et al., 2012; Loughran et al., 2014; Van de Schoot et al., 2021); however, its use within the field remains underexplored. Broader implementation has occurred in disciplines such as economics and marketing, where findings have been mixed

(Barrage & Lee, 2010; Bennett et al., 2018; Menapace & Raffaelli, 2020). The BTS has also been investigated in experimental philosophy, but early supportive findings have proven difficult to replicate (Schoenegger, 2021; Schoenegger & Verheyen, 2022). Recurring challenges include participants' scepticism about the method and doubts regarding promised bonus payments. This lack of trust may reduce engagement and undermine the mechanism's effectiveness. These challenges highlight the need for further research and tailoring the BTS mechanism to the psychological research context.

Current Study

This study aimed to evaluate the effectiveness of the BTS in improving the accuracy of self-report data in psychology, focusing on mitigating response biases related to sensitive questions. Given the potential challenges of participant scepticism about the mechanism and uncertainty regarding incentives (Barrage & Lee, 2010; Bennett et al., 2018; Menapace & Raffaelli, 2020), the study introduced an interim payment midway through the survey. This interim payment served a dual purpose: it aimed to demonstrate the researchers' ability to detect truthful responses and their commitment to fulfilling bonus payments. By integrating an interim payment as part of the BTS method, participants were expected to perceive both the mechanism and the incentives as more credible, thereby mitigating participant scepticism. In addressing these limitations, this study contributes to the literature by testing the BTS within a psychological context and evaluating a novel strategy to enhance its efficacy in sensitive self-report contexts.

Thesis Structure

This thesis follows a variation of the 'thesis by publication' approach and is structured around a research paper intended for publication. Specifically, the paper is a registered report, a format designed to improve research transparency and reproducibility through a two-stage peer review process. In Stage 1, the study's research questions, methodology and analysis plan undergo peer review before data collection. In Stage 2, after data collection and analysis, the full manuscript, including the approved protocol, results and discussion, is reviewed again. This approach ensures that studies are evaluated based on their theoretical and methodological rigour rather than the nature of their findings, helping to mitigate biases such as publication and reporting bias (Chambers, 2013).

This registered report was submitted to the Peer Community In Registered Reports (PCI RR), an initiative facilitating peer review for registered reports across disciplines. PCI RR does not publish articles but provides in-principle acceptance after a successful Stage 1 review. Accordingly, this study's research questions, methodology and analysis strategy were peer-reviewed before data collection, and in-principle acceptance was granted. The final manuscript was submitted after data collection and analysis, including the approved protocol, results and discussion. Once reviewers confirm adherence to the pre-registered plan and the validity of the conclusions, the report can be recommended by PCI RR and subsequently published in a PCI RR-friendly journal, which accepts and publishes PCI-recommended articles without further peer review (PCI, n.d.-a).

In light of this approach, this thesis comprises three main sections following this introduction:

- Literature Review: This section reviews the existing literature, providing the background and context for the research in more breadth than possible within the registered report due to word limits (see below). It establishes the study's theoretical framework and identifies gaps in the current literature that the registered report seeks to address.
- Registered Report: The registered report forms the core of this thesis. It includes the background, method, analysis, results, discussion and conclusion sections typically found in a peer-reviewed journal article. Once PCI RR approves the Stage 2 manuscript, it will be submitted to a PCI RR-friendly journal such as *Meta-Psychology* or *Advances in Methods and Practices in Psychological Science*. The latter journal has a 5,000-word limit, so the manuscript is, by necessity, concise and focused.
- Reflection: The thesis concludes with a reflection on the research journey, highlighting the challenges and benefits of adopting the registered report format. Situated within meta-psychology, this reflection explores key moments of academic and personal growth. Additionally, it serves as a potential resource for students and supervisors considering the registered report approach as part of a thesis.

Literature Review

(Un)Truthful Responding

Truthfulness is the cornerstone of reliable self-report data, yet participants do not always respond carefully or honestly. Cognitive effort and social pressures can influence responses, shaping the accuracy of self-reports (Paulhus & Vazire, 2007; Tourangeau & Yan, 2007). Given psychology's heavy reliance on self-report data (Baldwin, 2000), acknowledging these limitations and grappling with their implications is essential. Understanding the nature of (un)truthful responding is therefore critical, not only for evaluating the reliability of self-reports but also for recognising its broader impact on psychological research.

Self-Report Data

Self-report data commonly refers to information gathered through questionnaires or surveys typically self-administered in a written format, where respondents provide details about themselves. This method has a long history in the social sciences and has evolved with the advent of computerised surveys. Advances in technology and the rise of online survey platforms have further enhanced the convenience and affordability of self-reporting (Vecchio et al., 2020). Baldwin (2000) identifies two common scenarios for the use of self-report data. The first is when other data sources are challenging to obtain or not cost-effective. The second is when no alternative source is available, a situation frequently encountered in psychology.

For example, understanding internal or subjective constructs, such as self-esteem or attitudes towards socially sensitive issues, often relies on self-report data. Self-esteem represents an individual's assessment of their own value and worth,

primarily shaped by internal perceptions rather than external factors (Rosenberg, 1965). Attitudes towards socially sensitive issues may include opinions on topics like climate change, immigration, gender identity, racial equality and mental health stigma, to name but a few (Eagly & Chaiken, 1993). Research in these areas frequently uses self-reports to capture direct insights into people's thoughts and feelings, which may not be easily observable or measurable by other means.

To illustrate, alternative methods like physiological measures, behavioural observations and neuroimaging offer valuable insights but face significant limitations in fully capturing subjective experiences. For instance, physiological indicators like heart rate variability and cortisol can provide indirect information about stress or emotional arousal. However, while reflecting emotional regulation capacity, heart rate variability is influenced by non-emotional factors such as physical activity and breathing patterns, complicating its interpretation and limiting its ability to distinguish between emotional states (Appelhans & Luecken, 2006). Similarly, cortisol, often used as a biomarker for stress, is affected by factors such as circadian rhythms, diet and hormonal fluctuations. As a result, cortisol levels may not consistently reflect perceived stress due to the complex biological processes regulating its release and metabolism (Hellhammer et al., 2009; Miller et al., 2013).

Behavioural observations, such as facial expressions or body language, can provide additional context but may lack accuracy or clarity in interpretation. Emotional expressions can be consciously suppressed or exaggerated and are subject to cultural variability, making it difficult to draw universal conclusions about internal experiences (Mauss & Robinson, 2009). Furthermore, neuroimaging technologies, such as functional

magnetic resonance imaging (fMRI), provide insights into brain activity related to social decision-making. However, linking neural patterns to subjective experiences is inherently indirect and requires substantial interpretation (Ochsner & Lieberman, 2001; Poldrack, 2006). For example, fMRI might show activation in regions associated with emotional processing when discussing immigration, yet these activations cannot be directly translated into specific attitudes or beliefs.

Consequently, self-report data remains essential for directly assessing complex, deeply personal constructs. Unlike indirect methods, self-reports provide unique access to individuals' internal states and subjective experiences that cannot be fully captured through physiological, behavioural or neuroimaging proxies (Paulhus & Vazire, 2007; Podsakoff et al., 2003). However, the reliability of self-report data hinges on respondents' cognitive processes (Frank et al., 2017; Prelec, 2004).

Various biases can distort answers and hinder the accurate assessment of constructs, thereby impeding valid inferences (Alvarez et al., 2019; Arthur et al., 2021; Lilienfeld & Strother, 2020). The nature of self-report data—individuals' opinions, preferences, intentions and experiences—makes it particularly susceptible to social desirability bias (Chan, 2009). Therefore, while self-report data collection offers advantages in terms of ease and cost-effectiveness and may be the sole means of obtaining certain personal data, it simultaneously raises methodological and measurement concerns (Arthur et al., 2021). Given the reliance on self-report data in psychology, understanding and addressing response biases is critical to ensuring the accuracy of research findings. Among these biases, careless responding and SDR are

particularly impactful, with SDR being especially relevant in psychological research, where sensitive questions prone to bias abound.

Response Biases

Careless Responding. This category of response bias includes inattentiveness, poor response styles and even “an unwillingness to comply with testing demands” (Nichols et al., 1989, p. 240). With the rise of online self-administered surveys, concerns about response quality have grown. These platforms often lack supervision, thereby increasing the potential for increased careless responding (Frank et al., 2017; Vecchio et al., 2020). Even a small percentage of careless responders, estimated at 5-10% of a sample, can undermine the psychometric properties of a measure and compromise research findings (Arthur et al., 2021).

Inattentive responding, defined as providing answers without fully engaging with survey questions or tasks, is often characterised by careless reading and thoughtless responses (Krosnick, 1991; Meade & Craig, 2012). In extreme cases, respondents may guess randomly or choose the first acceptable response option, a behaviour known as satisficing (Berinsky et al., 2013; Oppenheimer et al., 2009). Response styles like extreme, midpoint or acquiescent responding (agreeing indiscriminately) also contribute to careless responding (Meade & Craig, 2012; Paulhus, 2002). Additionally, response patterns such as straight-lining (identical answers to consecutive items) and random responding undermine data integrity (Arthur et al., 2021; DeSimone et al., 2017).

Researchers use attention checks and response pattern indices, such as overall response duration, to identify careless responses (Huang et al., 2012; Maniaci & Rogge, 2014). Approaches to managing careless responses include dropping respondents

flagged as careless through attention checks or applying statistical adjustments (Alvarez et al., 2019; Vecchio et al., 2020). However, these methods have limitations. Excluding careless responders can reduce sample representativeness, and statistical adjustments can introduce new biases, complicating efforts to preserve data integrity (Arthur et al., 2021).

Socially Desirable Responding. Among response biases, SDR stands out as a significant force in shaping self-report data. It has been extensively studied in the social sciences (Fisher & Katz, 2000) and is frequently cited as a criticism of self-report data (Chan, 2009). SDR is not only a factor in how individuals present themselves or deceive themselves but also plays a crucial role in how they report their attitudes and behaviours (Simunović & Žeželj, 2023). Consequently, SDR has been shown to affect the measurement of personality variables (e.g., Mick, 1996), attitudes (e.g., Fisher, 1993) and self-reported behaviours (e.g., Mensch & Kandel, 1988). Like careless responding, SDR raises concerns regarding construct validity, impeding accurate assessments and valid inferences from data (Arthur et al., 2021). Fisher and Katz (2000) elaborate on the complex ramifications of SDR, which can attenuate, inflate or moderate relationships between variables (for a review, see Zerbe & Paulhus, 1987).

Unlike careless responding, SDR requires engagement with the item's content and entails portraying oneself positively to conform to societal norms (Paulhus, 1984; 2002). Paulhus (1984) identified two facets of SDR: intentional (impression management or faking) and unconscious (self-deception). Later, Paulhus (2002) expanded on this model, linking strategic agency (e.g., bragging) and self-deceptive enhancement to an inflated view of one’s social and intellectual status while associating

relationship management and self-deceptive denial to an exaggerated sense of one's moral qualities. SDR leads individuals to give responses that align more with perceived societal expectations than with their true feelings, beliefs or behaviours, resulting in the overreporting of socially desirable traits, attitudes and behaviours and the underreporting of undesirable ones (Arthur et al., 2021; Tourangeau & Yan, 2007).

Sensitive questions are especially likely to trigger SDR (Krumpal, 2013). A question is considered sensitive if it is perceived as intrusive by addressing 'taboo' topics that are typically regarded as inappropriate for everyday conversation (e.g. sexual behaviour) or if it involves a threat of disclosure, where respondents worry about the potential consequences of providing a truthful answer (e.g. admitting to illicit drug use). Additionally, a question is sensitive if it seeks a socially undesirable response, effectively asking the respondent to admit to violating a social norm (e.g. expressing a racist view). Social desirability concerns can be seen as a special case of the threat of disclosure, involving social disapproval as a specific type of interpersonal consequence of revealing information in a survey (Tourangeau & Yan, 2007).

Scales have been developed to detect social desirability, such as the 33-item Marlowe-Crowne Social Desirability Scale (MCSDS; Crown & Marlowe, 1960), in which the participant answers true or false to a set of socially desirable but improbable statements. These statements include "I have never deliberately said something that hurt someone's feelings." High scores indicate a greater need for social approval and a tendency to portray oneself positively. These scales may help researchers account for SDR by identifying individuals likely to present themselves in an overly favourable light.

However, respondents may manipulate their answers, complicating the detection of true SDR.

Strategies to mitigate response distortion effects include removal and statistical control. For instance, correcting the data of subjects with high social desirability scale scores (Nederhof, 1985) can involve using partial correlations or hierarchical stepwise regression analysis (Van de Mortel, 2008). However, similar to the challenges encountered with handling careless responding, these strategies may face difficulties in accurately identifying and quantifying the extent of SDR and risk inadvertently introducing bias through statistical adjustments. If SDR is measured with error, statistical control methods may not fully account for its influence unless they explicitly model measurement error in control variables (Westfall & Yarkoni, 2016). Based on the extant literature, Arthur et al. (2021) conclude that the statistical control of response distortion is not strongly recommended and should probably be avoided, noting that "when a respondent lies on a self-report measure, there is no way to convert the lies to truth; truth cannot be rediscovered statistically" (p. 122).

Research Method Implications

The paragraphs above emphasise the challenges associated with self-report data. Nevertheless, it remains indispensable in psychology, often providing the only viable method for accessing individuals' internal states and subjective experiences. The validity of psychological studies depends on the accuracy of the underlying data (Flake & Fried, 2020). However, biases such as careless responding and SDR persist as significant threats to the accuracy of self-report measures.

The call to address these challenges is not new. Baldwin (2000) highlighted the pressing need for behavioural researchers to strengthen data collection methods to address the inherent limitations of self-report data. More recently, Flake and Fried (2020) reiterated this need, arguing that researchers lack a solid foundation for drawing meaningful conclusions without credible evidence for the reliability and validity of measures.

One unequivocal way to enhance the reliability of self-report data is through the verifiability of responses (Becker & Colquitt, 1992; Cascio, 1975). In cases where responses can be objectively verified against an answer key, that is, a predefined set of correct responses used to measure accuracy, response biases can be effectively mitigated (Prelec, 2004). This approach ensures alignment between respondents' answers and objective benchmarks. However, this method is inherently limited to domains where such objective correctness is applicable. In psychological research, where private beliefs, attitudes and subjective states dominate, there is often no answer key; the ‘correct’ response is simply the one that most authentically reflects the respondent’s own perspective (Weaver & Prelec, 2013).

When verifiability is not possible, researchers often use post hoc strategies such as removal and statistical control to correct bias. However, these approaches are fraught with challenges. Risks include inadvertently introducing additional biases, reducing statistical power or mischaracterising the data (Arthur et al., 2021). Furthermore, these methods address biases only after they have distorted the data, meaning they are reactive rather than proactive solutions.

Therefore, the real challenge is developing methods that proactively minimise response biases. This challenge requires exploring ex-ante methods designed to reduce bias at the point of data collection. By aligning participants' motivations with attentive and honest responses, these methods should ensure that respondents provide the most accurate data possible, even when an answer key is not feasible.

Eliciting Truthfulness

Proactive approaches for mitigating response biases can be broadly categorised into psychologically grounded (non-incentive-based) and incentive-based. The former relies on cognitive and social strategies to encourage truthfulness, while the latter involves financial rewards designed to align participants' motivations with truthful responding. Psychologically grounded methods draw on theories such as social commitment theory (Cialdini, 1984) and goal-priming theory (Bargh et al., 2001). Incentive-based methods are grounded in Bayesian decision theory and mechanism design, which draw on principles from game theory to align individual incentives with accurate reporting. The distinction between financially driven methods and those purely reliant on social or cognitive factors highlights a prominent issue in contemporary psychology research practices. These approaches are explored further below.

Non-incentive-based Methods

Non-incentive-based methods leverage psychological strategies to encourage truthful responses and reduce reporting biases. Key techniques include participant anonymity, honesty priming, introspective beliefs, the solemn oath and cheap talk, each designed to elicit truthful responses to explicit questions. Additionally, the indirect

questions technique seeks to encourage honesty by indirectly addressing survey queries.

Participant anonymity protects the identities of study participants by ensuring that researchers cannot link responses back to individual participants. This method aims to achieve truthful responses by reducing fear of judgment (Podsakoff et al., 2003). However, while anonymity may decrease the motivation to distort reports in socially desirable directions (Joinson, 1999; but cf. Gnambs & Kaspar, 2017), it may only be effective in mitigating the impression management facet of SDR and not self-deception (Dwight & Feigelson, 2000). Furthermore, anonymity may diminish accountability, compromising measurement accuracy rather than improving it. For instance, three studies by Lelkes et al. (2012) demonstrated that allowing college students to answer questions anonymously sometimes increased reports of socially undesirable attributes but consistently reduced reporting accuracy and increased survey satisficing. These findings align with Weaver and Prelec's (2013) suggestion that anonymity techniques may not effectively address untruthfulness motivated by factors other than social sensitivity, such as boredom or carelessness.

Honesty priming involves procedures that implicitly encourage honesty and potentially careful responding. It draws on goal-priming theory (Bargh et al., 2001), which posits that activating specific goals, such as honesty, can shape subsequent behaviour. For example, honesty priming may involve asking participants to engage in tasks assessing the similarity of meanings of words related to honesty. Rasinski et al. (2005) found that college students exposed to words associated with honesty were more willing to admit engaging in socially undesirable drinking behaviours than those

exposed to unrelated words. However, goal-priming theory and related social priming methods have faced criticism regarding replicability and alternative explanations. Some high-profile priming effects have proven difficult to replicate (e.g., O'Donnell et al., 2018), raising concerns about their reliability. One key issue is demand characteristics, where participants adjust their responses based on perceived research expectations (Dalal & Hakel, 2016; Orne, 1962; Pashler et al., 2013). Another concern is experimenter expectancy effects, where researchers' unconscious cues influence participants' behaviour (Rosenthal, 1966). For instance, Doyen et al. (2012) attempted to replicate Bargh et al.'s (1996) study but found that experimenter expectations, rather than unconscious priming, likely explained the effect. These concerns indicate that honesty priming, like other social priming effects, requires further empirical validation to establish its robustness.

The introspective beliefs method uses prompts to encourage respondents to reflect on their intrinsic thoughts and values to increase self-report accuracy (Trautmann & van de Kuilen, 2015). While the main survey questions might focus on specific behaviours, for example, public transportation use, separate introspective questions are included to broaden the scope to personal values and beliefs, such as considering the importance of reducing carbon emissions from transportation in daily life. This approach encourages respondents to delve deeper into their perspectives, potentially leading to more truthful responding (Trautmann & van de Kuilen, 2015). However, this method's effectiveness relies on subjective introspective abilities (Pronin & Kugler, 2007). Furthermore, respondents may lack the motivation to think carefully about the questions, which requires more time and effort (Manski, 2004; Zizzio, 2010).

The solemn oath approach (Jacquemet et al., 2013) draws on social commitment theory (Cialdini, 1984), which posits that individuals are more likely to comply with requests if they feel a sense of commitment or obligation to a social group or cause. This approach aims to foster truthful answers by invoking a sense of moral or ethical duty in participants by signing an oath or commitment, such as pledging to provide truthful responses. However, the effectiveness of the solemn oath remains uncertain, with recent willingness to pay studies (e.g., Mamkhezri et al., 2020) yielding mixed results. These economic studies are pertinent in psychology as they highlight the potential for bias when individuals align their responses with social norms or expectations rather than genuine preferences or behaviours. The efficacy of the solemn oath may depend on the seriousness with which participants approach the commitment, indicating that individual characteristics might play a crucial role in its effectiveness.

Cheap talk refers to providing instructions that stress the importance of giving truthful responses. The term originates from game theory, where truth-telling can occur through partial alignment of interests between two parties (Farrell & Rabin, 1996). In survey research, Cummings and Taylor (1999) formalised using cheap talk to mitigate hypothetical bias in contingent valuation studies (see the subsection Experimental Applications – Economics for a description of contingent valuation). They provided scripts explaining these biases and instructed respondents to answer questions as if in a real-world situation. Although Cummings and Taylor (1999) found that cheap talk effectively eliminated hypothetical bias, later studies reported mixed results (Ladenburg & Olsen, 2014). The key limitation of cheap talk in surveys is that no mutually aligned interest exists, as truth-telling primarily benefits the researcher while imposing a mental

cost on subjects who must respond more carefully. Consequently, there is no incentive for subjects to answer survey questions sincerely, making this approach reliant on the goodwill of respondents. Due to this structural limitation, the cheap talk approach might not work when a survey question requires a higher mental cost or when there is a motive for providing deceptive responses, such as concerns about self-image (Lee, 2023).

The indirect questions technique (Campbell, 1950) involves asking respondents to provide answers from another person's or generalised other's perspective. Accordingly, it aims to enhance respondent comfort and anonymity to encourage more honest disclosures. Fisher (1993) and Fisher and Tellis (1998) found that indirect questioning reduced both types of SDR (impression management and self-deception) in sensitive topics, revealing true attitudes and beliefs that direct questioning might obscure. However, by its very design, this method does not address explicit questions directly, which introduces complexity in question design and data interpretation. Additionally, it does not address careless responding.

Despite the variety of bias-mitigating approaches discussed above, no single approach comprehensively addresses all facets of response bias. Their effectiveness is often context-specific and constrained by inherent limitations (Lee, 2023). A recurring theme is the dependence on individual differences. The introspective beliefs and solemn oath techniques rely on participants' ability for introspection or willingness to adhere to a moral commitment, characteristics that can vary widely across individuals. Another key challenge lies in the trade-off between reducing bias and maintaining accountability. For example, participant anonymity effectively minimises impression

management but may compromise accuracy by reducing accountability, while indirect questions address self-deception yet fail to mitigate careless responding. Moreover, honesty priming and cheap talk may introduce new vulnerabilities, including experimenter demand effects and reliance on participants' goodwill. The complexity of some techniques, like indirect questions, further limits their applicability, particularly when explicit or straightforward questioning is required.

A potentially critical limitation of many psychologically grounded methods is their reliance on intrinsic motivations, which may not always be sufficient to elicit truthful responses (Crowne & Marlowe, 1964; Deci & Ryan, 1985). Intrinsic motivation refers to engaging in a behaviour because it is inherently satisfying or aligned with one's values and sense of self. Methods such as honesty priming, introspective beliefs, the solemn oath and cheap talk rely heavily on participants' intrinsic willingness to reflect, commit to moral obligations or overcome cognitive effort. For instance, the solemn oath depends on individuals' willingness to honour a moral commitment, which may vary widely across cultural or personal contexts. Consequently, these methods can fall short when intrinsic motivation is weak, such as when questions are deeply sensitive, cognitively taxing or fail to resonate with participants' values. In such cases, extrinsic motivators, which involve engaging in behaviour to obtain external rewards (Deci & Ryan, 1985), may provide a complementary solution. Techniques like incentive-based methods harness extrinsic motivation by aligning participants' self-interest with truthfulness.

Incentive-Compatible Methods

Incentive-based methods that provide direct financial rewards for truthfulness address what Schoenegger (2021) calls the 'incentivisation challenge' in psychology

and related disciplines. In the social sciences, compensation for surveys often rewards task completion rather than response quality, exacerbating the problem where intrinsic motivation alone is insufficient to overcome competing pressures. This concern is particularly relevant with the increasing use of online, unsupervised surveys in psychological research. In this context, participants may be more likely to provide superficial or careless answers, knowing they are likely to be paid regardless of response accuracy (Peer et al., 2021). To overcome this limitation, incentive-based methods explicitly tie compensation to truthful responses, reflecting both attentiveness and accuracy, aligning participants' self-interest with researchers' objectives. Drawing on principles from economics, such as the principal-agent problem (Jensen & Meckling, 1976), these methods aim to address the cognitive, temporal and emotional costs of truthfulness, encouraging more deliberate and honest engagement.

Incentive-compatible methods operationalise these principles by designing systems where truthful responses become the most rewarding strategy for participants (Toulis et al., 2015; Weaver & Prelec, 2013). Unlike traditional incentive-based approaches that merely reward participation, incentive-compatible mechanisms directly structure rewards to favour truthfulness. For example, participants achieve the highest payoffs when their responses align with truthful reporting, ensuring that truthfulness is not only encouraged but also incentivised. This shift from reactive adjustments to proactive systems represents a significant advancement in addressing the incentivisation challenge.

Incentive-compatible mechanisms often use scoring rules to assess the quality of participants' responses. In subjective contexts, scoring rules serve a role akin to answer

keys in objective settings by providing a structured framework that incentivises truthful answers. These rules assign scores based on how well participants' responses align with expected or presumed truthful outcomes. Therefore, scoring rules represent a proactive attempt to achieve a level of reliability comparable to that of verifiable situations, thereby addressing the challenges of bias when direct verification is not possible. The specific design of the scoring function depends on the mechanism in use and its underlying assumptions.

For instance, certain incentive-compatible mechanisms score and reward participants based on the prevalence of their answers within the entire sample. This is based on the idea that if many people give the same answer, it is more likely to be correct or truthful. Therefore, participants are incentivised to select answers that align with those chosen most frequently by other participants. However, this approach risks incentivising conformity rather than genuine responses, as participants may prefer to guess what they believe will be the most common answer instead of expressing their true thoughts or judgments (Prelec, 2004).

An enhancement to these methods is the classical peer prediction technique proposed by Miller et al. (2005). This method assesses participants' responses relative to their peers and determines scores or payments accordingly. It relies on a positive correlation between an individual's own experiences and the experiences of others. Even when participants perceive their experiences as potentially differing from others, the effectiveness of peer prediction still depends on this positive correlation. For example, a participant who has used illegal drugs might estimate a 40% likelihood of others having used them as well, whereas a participant who has not used drugs may

give a much lower estimate of 20% (Witkowski, 2014). In this way, individuals' responses are implicitly shaped by their own experiences or beliefs, which serve as a 'sample of one' to guide their predictions.

However, the peer prediction method is not without its limitations (Schoenegger, 2021). A significant drawback is its dependence on participants' shared prior beliefs and common knowledge (Witkowski, 2014). In Bayesian theory, a prior belief represents the initial subjective probability assigned to a hypothesis before incorporating new data or evidence (Gelman et al., 2014). These priors are updated in light of new evidence using Bayes' theorem. The classical peer prediction method assumes that all participants' prior beliefs are identical and known and must be accounted for to compute the scoring or payment rule accurately. However, this assumption can be impractical in real-world applications, where prior beliefs are often heterogeneous. This reliance on shared priors limits the method's applicability, particularly in diverse populations where prior beliefs may vary significantly.

The Bayesian Truth Serum (BTS), introduced by Drazen Prelec in 2004, represents a significant advancement over traditional scoring techniques in several ways. Like other Bayesian mechanisms, the BTS leverages the correlation between a person’s opinion and the opinions of others. This builds on the principle that an individual’s own experiences or beliefs are informative about others, as described in more detail in the next chapter. However, unlike previous methods, the BTS does not incorporate assumptions about this correlation into its scoring function. This departure from relying on predetermined probabilities marks a substantial improvement compared to the peer prediction approach (Schoenegger, 2021; Witkowksi, 2014). Specifically, the

BTS removes the need for questions to have empirically estimated prior and conditional probabilities that are central to the peer prediction method's limitations. By eliminating this requirement, the BTS offers greater flexibility and applicability across different populations, making it particularly promising for addressing the incentivisation challenges in psychology and related disciplines (Schoenegger, 2021).

Bayesian Truth Serum

The BTS (Prelec, 2004) offers a quantitative method for encouraging truthful (considered and honest) responses to subjective questions by scoring the truthfulness of responses and rewarding higher scores with a bonus payment. The approach leverages a Bayesian interpretation of population distributions, aligned with the false consensus effect, whereby individuals typically overestimate the prevalence of their own opinions, assuming their views are more common than they actually are (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’. The BTS ties incentives (bonus payments) to these surprisingly common responses to motivate participants to provide truthful answers rather than being careless, conforming to social desirability or guessing what others might say.

By integrating Bayesian principles and game theory, the BTS seeks to establish a framework where truth-telling is participants' optimal and self-serving choice. Theoretically, this is achieved by creating a Bayesian Nash equilibrium in which truth-telling maximises a participant’s chance of achieving a higher score. In this equilibrium, truth-telling is stable because participants cannot improve their expected

payoffs by deviating from this strategy, given that others are also assumed to be truthful (Prelec, 2004; Witkowski, 2014).

BTS Operation

The BTS operationalises the alignment of self-interest with truthful reporting through thoughtfully designed scoring rules. The method begins by advising participants that the survey uses an algorithm developed to encourage truth-telling. Participants are informed that this algorithm will assign numerical scores based on how truthful their answers are and that those with the highest scores will receive a bonus on top of their base pay. Figure 1 shows the standard BTS text provided to participants.

Participants then proceed through two stages when answering survey questions. First, they provide their personal response to each question; then, they predict the proportion of other respondents who will select each response option. For example, in a question asking, "Do you like dogs?" participants not only indicate whether they like dogs but also estimate the percentage of other respondents who would answer 'yes.' While participants are typically not told exactly how the scoring works, the BTS operates on the basis that they understand and trust that giving thoughtful and truthful responses will increase their chances of earning a bonus. At the end of the study, participants receive a base payment, with additional bonuses awarded to those whose responses are deemed more truthful based on the BTS scoring criteria.

Theoretical Foundation

Bayes' theorem, developed by the Reverend Thomas Bayes in the eighteenth century, forms the foundation of the BTS. As a fundamental concept in probability

theory, it guides how the probability of an event is updated in light of new evidence, using conditional probabilities to refine understanding in uncertain situations.

The theorem is grounded in the joint probability of two events, A and B, which can be expressed as:

$$P(AB) = P(A|B) * P(B) = P(B|A) * P(A)$$

By rearranging the above equation, Bayes Theorem emerges as follows:

$$P(A|B) = (P(B|A) * P(A)) / P(B)$$

Where:

- P(A|B) represents the probability of event A occurring, given that event B has occurred, known as the posterior probability.
- P(B|A) represents the probability of event B occurring, given that event A has occurred, referred to as the likelihood.
- P(A) denotes the probability of event A occurring, termed the prior probability.
- P(B) signifies the probability of event B occurring, known as the marginal likelihood or evidence.

In Bayesian probability theory, one of these events typically represents the hypothesis, H, and the other is the data, D. The goal is to assess the relative truth of the hypothesis given the data, and Bayes' rule expresses this relationship as:

$$P(H|D) = (P(D|H) * P(H)) / P(D)$$

Here, P(H|D) denotes the posterior probability, representing the updated belief in the hypothesis after considering the data. The likelihood function, P(D|H), represents

the probability of observing the data under the assumption that the hypothesis is true, typically reflecting the researcher's expectations about the data under the assumption of the hypothesis being true. The prior probability, P(H), reflects the initial belief in the hypothesis before considering the data and is often considered the most subjective aspect of Bayesian probability theory. Finally, P(D) represents the overall probability of observing the data. As calculating P(D) independently can be challenging, Bayes' rule is often expressed proportionally as:

$$P(H|D) \propto (P(D|H) * P(H))$$

Bayes' theorem explains how beliefs are updated with new information, blending prior knowledge with fresh data to derive updated probabilities. This theoretical foundation underpins the practical implementation of the BTS. As described previously, participants provide both their personal responses to a survey question and a prediction of how others will respond. The prior reflects the participant’s belief about the distribution of responses in the population before considering their own answer. While this belief may not be consciously formulated, it serves as a baseline expectation of how others will respond. The posterior is the participant’s updated belief about the population distribution after incorporating their own response, informing the predictions they enter regarding how others will respond.

To illustrate, consider Anne, who answers 'Yes' to 'Do you like dogs?' Before considering her own response, her prior is her general belief about how common 'Yes' and 'No' responses are in the population. Once she answers 'Yes,' she updates her belief, forming a posterior that integrates this new information. Based on this updated belief, she predicts that around 70% of respondents will also answer 'Yes.' Other

participants may predict different distributions depending on their own responses. For example, someone who does not like dogs might predict a lower percentage, such as 50%. Since the group collectively underestimates how common 'Yes' responses are, the BTS scoring mechanism rewards Anne for reporting a surprisingly common response. By linking incentives to the relationship between predicted and actual response distributions, the BTS encourages participants to provide truthful answers.

Calculation of Truthfulness (BTS) Scores

The BTS functions at the level of an individual question, assigning a truthfulness score (subsequently referred to as the ‘BTS score’) to each answer. This 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.

Information Score. 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.

Prediction Accuracy Score. 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.

Combined BTS Score. 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:

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

Rationality of Truth-Telling within BTS

Truth-telling in the BTS framework is considered individually rational because the scoring system is designed to maximise each participant's utility when they provide truthful responses. Utility refers to the subjective value participants assign to outcomes like earning a bonus for truthful responses relative to strategies like deception. This principle aligns with Rational Choice Theory (RCT), which posits that individuals make decisions to maximise utility based on preferences, constraints and available information (von Neumann & Morgenstern, 1944). By linking self-interest to truthfulness, the BTS creates conditions where truth-telling maximises expected payoffs, reflecting

principles of Expected Utility Theory (EUT; Bernoulli, 1738/1954). In this Bayesian Nash equilibrium, truth-telling becomes the dominant strategy, meaning that it is the strategy that results in the highest payoff for a participant, regardless of what others do. In other words, truth-telling is the best choice for a participant because it maximises their expected outcome, and no participant can improve their payoff by choosing a different strategy, assuming others are also truthful (Nash, 1950; Prelec, 2004).

In real-world scenarios, however, individuals may not consistently exhibit the behaviour expected of rational agents (Trautmann & van de Kuilen, 2011; Weaver & Prelec, 2013). While the intrinsic rationality of truth-telling is essential for the BTS to elicit genuine responses effectively, practical conditions may diverge from the assumptions of rationality and Bayesian behaviour. The BTS theory posits that individuals iteratively update their beliefs or predictions using Bayes' rule, starting with a common prior belief (prior distribution) regarding the population's responses and subsequently adjusting this belief in the direction of their personal responses. However, the notion of a 'prior' belief, devoid of personal preferences, may lack psychological realism (Weaver & Prelec, 2013). In practice, individuals' prior beliefs are likely to be shaped by subjective experiences, biases and cultural influences, which can diverge from the theoretically objective 'common prior' assumed by the BTS (Trautmann & van de Kuilen, 2011). Furthermore, predictions made by respondents with similar mindsets may exhibit significant variability in practice owing to differences in individual experiences, knowledge and interpretation of the question.

These potential limitations of the BTS highlight the importance of empirical validation. Examining how well the mechanism performs in practice, particularly in

contexts where participants' decision-making may deviate from theoretical expectations, is crucial to understanding its practical effectiveness. Identifying the specific conditions that enable or hinder the BTS could help refine its application and address challenges in real-world implementations.

BTS Validation: Empirical Studies & Practical Applications

While the BTS's theoretical principles provide a robust foundation for incentivising truth-telling, its practical efficacy must be evaluated empirically to address real-world complexities and participant behaviours. This need is especially pronounced in psychology, where SDR poses particular challenges to the validity of self-report data. The unique sensitivity of psychological constructs necessitates targeted empirical research to assess how effectively the BTS can reduce bias and enhance the accuracy of self-reported data.

Although the BTS's applications in psychology remain limited, insights from related disciplines, such as criminology, experimental philosophy, economics and marketing, provide valuable context. In economics, contingent valuation studies reveal parallels between hypothetical bias and SDR. Moreover, marketing applications demonstrate the BTS’s potential to predict behavioural intentions. These cross-disciplinary findings offer a foundation for refining the BTS in psychology.

Validation Studies

The largest validation study of the BTS, conducted by Frank et al. (2017), though outside the field of psychology, offers compelling foundational evidence supporting the method's efficacy in incentivising truthful responses. These large-scale online human experiments comprised two scenarios involving stochastic tasks, i.e., tasks with random

outcomes, specifically coin flips and dice rolls. Additionally, the BTS was applied in a more applied context through a realistic pricing survey to demonstrate its practical utility.

In the coin flip experiment, participants were randomly assigned to either a BTS treatment or a control group. They flipped a coin five times, reported the outcomes and received a bonus for reporting heads. Participants were also asked to predict the proportion of heads reported by all participants. The BTS treatment increased the proportion of reported tails from 43% in the control group to 47%, moving closer to the expected 50% in an honest reporting scenario. This improvement in honesty was statistically significant according to the binomial test, suggesting that the BTS can effectively reduce the tendency to exaggerate outcomes in scenarios where responses can be easily manipulated.

The dice experiment provided a more complex context for evaluating the BTS. Participants were randomly assigned to a control group or one of two BTS treatments, one with a dynamic view of their i-score and one without. Participants rolled a dice five times, reporting the outcomes, with rewards based on the sum of the reported dice rolls at a rate of $0.01 per point on the sum of the dice. The experiment allowed for a broader range of dishonest behaviours, such as claiming all sixes to maximise rewards or reporting plausible combinations of high numbers. Both BTS treatments led to statistically significant improvements in honesty over the control group, as indicated by Pearson's χ^2 goodness-of-fit test. However, even within the BTS conditions, the distributions of reported outcomes showed a rightward skew, suggesting a bias toward profit maximisation, albeit less pronounced than in the control group. Interestingly, the BTS condition where dynamically calculated i-scores were visible to participants did not

significantly improve over the BTS condition without visibility of i-scores, suggesting that simply knowing the BTS scoring criteria was sufficient to encourage more honest behaviour in this context.

In their pricing experiment, Frank et al. extended testing of the BTS to a decision-making under uncertainty setting. Participants were presented with a sample question and asked to choose a reward for completing a multiple-choice questionnaire of 20 such questions. Reward options ranged from $0.10 to $1.00 in 10-cent increments. Participants were randomly assigned to conditions similar to the dice experiment (control, BTS with dynamic i-score visibility, and BTS without it). The control group tended to select the maximum reward ($1) for completing the questionnaire. In contrast, the proportion of participants choosing this option decreased in the BTS conditions, as determined by Pearson's χ^2 test. This outcome suggests that the BTS incentivised more realistic self-assessments regarding compensation. The finding that the BTS with dynamic i-score visibility yielded significantly different response patterns than other conditions points to the utility of i-score transparency in increasing participants' awareness of response quality.

These experiments by Frank et al. (2017) demonstrate that the BTS can reduce bias and improve the truthfulness of self-reports across various tasks. The findings also confirmed a fundamental assumption of the BTS methodology: whether individuals regard their personal opinions as indicative signals of the broader population distribution. Significant correlations between participants' predictions and their endorsed responses across all conditions supported the idea that people tend to expect a higher probability of their own opinions aligning with the group consensus. This

assumption was also validated in related research by Yang et al. (2016). It found that participants predicting specific outcomes (e.g., NBA playoff victories) were more likely to estimate a higher probability of the group's predictions matching their own.

Experimental Applications

Economics. The BTS has mainly been applied in contingent valuation studies and soliciting expert opinions within economics. While expert opinion studies, such as Yang et al. (2016) and Zhou et al. (2019), show promising alignments between BTS forecasts and actual outcomes, the results are less conclusive in contingent valuation contexts. Contingent valuation involves presenting respondents with hypothetical scenarios and asking about their willingness to pay for specific benefits or accept compensation for losses (Carson, 2012; Mitchell & Carson, 1989). A key challenge in such studies is hypothetical bias, the tendency for individuals to overstate their willingness to pay or behave differently in hypothetical scenarios compared to real-world situations (Cummings et al., 1995). In this context, truthful responses about willingness to pay should correspond closely to actual purchasing behaviour, referred to as revealed preference data.

Barrage and Lee (2010) examined the effectiveness of the BTS compared to other truth-eliciting methods by assessing donation decisions for public goods across various treatments (Real, Hypothetical, Cheap Talk, Consequentialism and BTS). The public goods included providing tents for disaster relief or contributing to staff pay for operating a pollution victims hotline. As expected, significantly different responses were observed between real and hypothetical conditions ($p < .05$), with greater willingness to donate in the hypothetical scenario. However, the BTS exhibited inconsistent success: it

reduced but did not wholly eradicate hypothetical bias for one good (the hotline) and had little effect in reducing bias in the case of the other (tents). The authors speculated that participants' unfamiliarity with the BTS may have fostered mistrust, limiting its ability to incentivise truthful responses.

This interpretation of findings aligns with those of Bennett et al. (2018), who failed to detect a significant difference between responses with and without BTS incentives in a study investigating whether people were willing to pay extra for the provisions of the EU Broiler Directive, which set minimum standards for the protection of chickens raised for meat production. Debriefing questions revealed that 59% of participants in the BTS group agreed with the statement that they believed the information provided by the researchers about the BTS, indicating that 41% did not. The researchers also considered whether the incentive for truth-telling was too small and uncertain. In their study, a winner was randomly selected from the BTS condition to receive a prize of up to £100, with the exact amount awarded based on their truthfulness score. Furthermore, Bennett et al. conjectured that the absence of significant differences in willingness to pay between non-incentivised and BTS-incentivised groups may have been due to the perceived gravity of the issue, which may have reduced the scope for differences between the two conditions.

Lee (2023) expanded on Barrage and Lee’s (2010) research by focusing on the idea that the specific context of the donation target could be instrumental. Participants were asked to consider donating to 'children suffering from cancer', a scenario expected to elicit empathy. Despite partial mitigation of hypothetical bias by the BTS (61.3% versus 45.3% chose to donate in the hypothetical and BTS conditions, respectively), a

significant gap remained between responses in the BTS condition and those in the real condition, where only 29.7% chose to donate. Using the two-sample test of proportions, the responses in the real group were significantly different from those in the BTS group ($p < .05$).

Menapace and Raffaelli (2020) reported similar findings in a direct choice experiment involving grocery purchases. The researchers measured the proportion of overlap between the probability distributions of willingness to pay responses in the BTS treatment and actual purchasing behaviour. They compared this with the overlap observed in the control condition. Using the overlapping coefficient as their metric, they found that the BTS condition produced a higher overlap (0.51 across six attributes) than the control condition (0.23). However, bias was not fully eradicated, as a coefficient of 1 would indicate perfect alignment. Additionally, the study showed that the BTS reduced willingness to pay for product attributes with a public component (i.e., those benefiting society) but had no effect on attributes with a private component (i.e., those benefiting individual consumers).

Marketing. Howie et al. (2011) used the BTS in a marketing context to investigate forecasts for new product adoption by administering a survey before the product launch to gauge respondents' intentions. This approach is similar to contingent valuation studies, as respondents also make decisions in hypothetical scenarios. In the Howie et al. study, a comparison of respondents' adoption intentions with actual adoption rates after the product launch revealed that the BTS scenarios outperformed a reference model, improving predictive accuracy by up to 36%.

In a frequently cited study in the BTS literature, Weaver and Prelec (2013) demonstrated the BTS's efficacy in reducing biases in recognition questionnaires containing fictitious items where respondents might falsely claim recognition due to social desirability or perceived pressure to appear knowledgeable. This study found that participants in the BTS condition exhibited a reduced recognition of non-existent items compared to control groups, even when presented with conflicting incentives to overstate their responses. Notably, individuals who received BTS-based payments without explicit instructions displayed a decreased likelihood of recognising fictitious items over time, indicating a learned response to BTS incentives. Moreover, when applied to judgments regarding contributing to a public good, the BTS eradicated biases commonly observed in contingent valuation solicitations.

Psychology and Related Disciplines. Applications of the BTS in psychological research showcase the method's potential to elicit truthful responses in sensitive contexts. For instance, the BTS was used to examine honesty in reporting questionable research practices (QRPs) within academic settings (John et al., 2012; Van de Schoot et al., 2021). Additionally, its use in criminology has revealed variations in individuals' self-reported offending behaviours and intentions to offend (Loughran et al., 2014). These examples illustrate the potential of the BTS to address challenges in obtaining accurate self-reports on socially sensitive issues.

John et al. (2012) surveyed over 2,000 psychologists, investigating personal engagement in QRPs and perceived prevalence and admission among peers. Respondents in the BTS condition were told that the researchers would donate to a charity of their choice (selected from five options) and that the overall donation would

depend on the truthfulness of their responses. In the control condition, participants were simply informed that a donation would be made on behalf of each respondent to a maximum combined total of $2,000. Across both conditions, around $4,200 was donated to the charities. The researchers hypothesised that the BTS condition would influence responses relative to baseline levels in the control condition. While these baseline levels were unknown, they expected the effects to be minimal for impersonal estimates of prevalence and admission and greatest for personal admissions of unethical practices. Results generally aligned with these expectations. High odds ratios were observed for self-admission rates in the BTS condition compared to the control condition for certain practices like falsifying data (OR = 2.75). In contrast, moderate odds ratios were observed for items such as prematurely stopping data collection (OR = 1.57), falsely claiming that results are unaffected by certain variables (OR = 1.52), and falsely reporting a finding as expected (OR = 1.45).

In Loughran et al.'s (2014) study, young adults were presented with four hypothetical anti-social scenarios: drinking and driving, smoking marijuana, cheating on an exam and texting while driving. The study uncovered significant differences in responses between the BTS incentive condition and the control condition for drunk driving and cheating on an exam. For instance, in the drunk driving scenario, the mean reported willingness to offend was approximately 16 percentage points higher in the BTS condition (53.2%) than in the control condition (36.8%; $t = 3.04$, $p = .003$). However, there were no significant differences observed for marijuana usage and texting while driving, which was thought to be due to their greater social acceptability among the demographic.

In his 2021 empirical study, Schoenegger applied the BTS to seven vignettes from recent philosophical psychology journal articles. The vignettes covered diverse topics such as attributions of knowledge-how, freedom of choice in nudging scenarios and the moral permissibility of torture. Participants were randomly assigned to either a control or BTS treatment condition, where they read the vignettes and responded using a seven-point Likert scale. To incentivise truthful responses, participants in the BTS condition were informed that an additional £1 bonus would be awarded to the top third of scores. Pearson's χ^2 tests revealed significant differences ($p < .001$) in response distributions between the BTS and control groups for four of the seven vignettes, specifically those concerning knowledge-how, freedom of choice, correspondence theory of truth and determinism. These results suggest that the BTS can influence response patterns in philosophical contexts, potentially promoting more thoughtful or honest reporting.

In a subsequent 2022 study, Schoenegger and Verheyen aimed to replicate and extend these findings, conducting the first registered report application of the BTS. The study sought to reproduce the results of the earlier research and discern any unique effects of the BTS. They introduced two additional conditions: the "Additional Money" and the "Prediction" conditions. In the "Additional Money" condition, participants were offered monetary rewards equivalent to those in the BTS treatment but without the requirement to provide predictions, ensuring that any observed differences in the BTS condition could not simply be attributed to financial incentives. In the "Prediction" condition, participants made predictions about others' responses but received no bonus

for truthfulness, allowing the researchers to examine the effect of prediction-making independently of incentive structures.

The participant pool was distributed across four conditions: Regular Incentive, Additional Money, Prediction and BTS. The primary pre-registered analyses did not reveal significant differences between the BTS and the Regular Incentive (control) condition for any of the seven items, indicating a replication failure based on the pre-registered criteria. However, “weak evidence” emerged in comparisons involving the Additional Money and Prediction conditions versus the BTS, with two items significantly differing between the Additional Money and BTS conditions and one item significantly differing between the Prediction and BTS conditions. In non-pre-registered analyses, “weak evidence” also indicated significant differences between the Prediction condition and the control (one significant item). In contrast, no significant effects were observed when comparing the Additional Money condition to the control. Given the study's limited power to detect all potential effects, the authors refrained from making definitive conclusions about the broader applicability of the BTS mechanism in social science research and advocated for further investigation.

Summary of Empirical Findings

The BTS literature reveals diverse outcomes across disciplines, ranging from significant improvements in truthful responding to instances where bias mitigation was minimal or absent. Seminal studies, such as those by Frank et al. (2017), John et al. (2012) and Weaver and Prelec (2013), demonstrate the BTS’s ability to reduce SDR and enhance self-report accuracy. Frank et al. highlighted the effectiveness of the BTS in incentivising honest reporting in tasks like coin flips and pricing scenarios, while John

et al. showcased its utility in sensitive domains like research ethics. Similarly, Weaver and Prelec successfully reduced over-claiming in recognition tasks, illustrating the BTS's adaptability across varied contexts.

In contrast, studies such as Barrage and Lee (2010) and Bennett et al. (2018) report mixed or null findings, highlighting the challenges of consistently applying the BTS. For instance, in contingent valuation contexts, Bennett et al. identified participant scepticism about the BTS and doubts regarding the incentive structure as critical barriers to effectiveness. Likewise, Schoenegger and Verheyen's (2022) replication attempt failed to provide evidence supporting the BTS. These findings raise questions about the robustness of the BTS across different populations and study designs.

Trust emerges as a critical factor influencing the BTS's efficacy. It can be conceptualised as a willingness to accept vulnerability based on positive expectations of another's intentions, a concept crucial for cooperative tasks (Rousseau et al., 1998). In the context of the BTS, participants may feel vulnerable to risks such as unfulfilled promises, which can reduce engagement and undermine the mechanism's effectiveness. The empirical findings suggest that a lack of trust is a recurring challenge in BTS applications. For example, Menapace and Raffaelli's (2020) trust test yielded mixed results, demonstrating participants' hesitancy to trust BTS incentives fully. This scepticism reflects broader challenges in fostering trust in a mechanism that operates without explicit explanation. Furthermore, several studies suggest that insufficient or uncertain incentives for truth-telling can undermine the BTS (Bennett et al., 2018; Menapace & Raffaelli, 2020). Doubts about receiving bonus payments in general, common among online survey participants, can exacerbate these challenges. These

findings align with psychological theories of trust, emphasising the importance of credibility, predictability and transparency in fostering cooperation and honesty (Mayer et al., 1995; Rousseau et al., 1998).

Frank et al.'s (2017) findings highlight the role of transparency in addressing trust issues. Transparency fosters trust by addressing participants' need for predictability and reliability in the incentive structure (Mayer et al., 1995). In Frank et al.'s pricing experiment, providing participants with dynamic i-score visibility enhanced honesty by helping them better understand how their responses were evaluated. However, the feasibility of such transparency in subjective self-reports with no clear 'right' answers remains limited. This highlights a broader gap in the literature: while trust is recognised as central to the BTS's effectiveness, empirical work exploring interventions explicitly designed to enhance trust, such as explicit demonstrations of incentive reliability, is minimal. Furthermore, studies testing the BTS in psychologically sensitive contexts, where the interplay between trust and SDR may be particularly pronounced, are notably lacking.

Current Study

Building on the insights and limitations identified in the existing literature, the current study aimed to enhance the applicability of the BTS methodology in psychology by addressing critical gaps related to participant trust. Specifically, it evaluated the BTS's effectiveness in improving the accuracy of self-reported data by mitigating biased responses to sensitive questions. These questions, drawn from established scales measuring social dominance, sexism and ageism, were selected for their propensity to provoke social desirability bias. While SDR was the primary focus, the study adopted

Prelec's dual definition of truthfulness, encompassing both careful consideration and honesty. It was anticipated that the BTS would encourage more thoughtful responses, with the built-in benefit of reducing careless responding due to its engagement-focused design. To address participant scepticism regarding the mechanism and the payment of incentives, the study introduced an interim payment midway through the survey. This approach aimed to demonstrate the researchers' ability to detect truthfulness and commitment to fulfilling promised incentives. By combining sensitive questions with trust-building measures, the study sought to enhance participants' confidence in the BTS and, in turn, improve the accuracy of self-reported data in contexts where SDR is most pronounced.

Registered Report

Preamble

The following section comprises the research paper prepared as a registered report with the PCI RR. This initiative promotes transparent and reproducible science by reviewing and recommending preprints through a two-stage peer review process (PCI, n.d.-a). Accordingly, this study's questions, hypotheses, methodology and analysis plan were peer-reviewed before data collection. Following a detailed review and revision, the proposal received in-principle acceptance. After completing the research, a manuscript containing the approved protocol, results and discussion was submitted. At the time of writing, the Stage 2 manuscript is being assessed for compliance with the protocol and the validity of conclusions. Provided the manuscript is positively recommended, the study will be eligible for publication in PCI RR-friendly journals without further peer review.

The introductory section of the registered report, reproduced below without any changes, draws on the findings from the literature review. Consequently, there is some overlap with the preceding sections of this thesis. This overlap is due to the necessity of including an appropriate background in both the thesis (in the form of the comprehensive literature review) and the registered report (in the form of a condensed summary of the extant literature and need for the study), as well as the requirement to keep the registered report unchanged after final peer review. The subsequent sections of the registered report then detail the study hypotheses, method, analysis, results, discussion and conclusion.

Claire M. Neville and Matt N. Williams co-authored the registered report. The contributions of each author are outlined below using the CRediT (Contributor Roles Taxonomy) framework (Brand et al., 2015):

- Claire M. Neville: Conceptualisation, Methodology, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualisation, Project Administration, Funding Acquisition.
- Matt N. Williams: Conceptualisation (supporting), Methodology (supporting), Writing – Review & Editing, Supervision, Funding Acquisition.

Full Reference for the registered report pre-print is as follows:

Neville, C. M., & Williams, M. N. (2025). *Does truth pay? Investigating the effectiveness of the Bayesian truth serum with an interim payment: A registered report.* OSF Preprints. https://doi.org/10.31219/osf.io/s3znc

Research Paper: 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 [2]

Self-report data is vital in psychological research, but biases like careless responding and socially desirable responding (SDR) 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. However, the effectiveness of the BTS across disciplines remains inconclusive, with possible challenges including participant disbelief and uncertainty regarding incentives. This study evaluated the effectiveness of the BTS in mitigating SDR 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, response patterns diverged from the mechanism's intended effects, raising concerns

[1] School of Psychology, Massey University

[2] This abstract is largely similar to the thesis abstract but is more concise, focusing strictly on the study's rationale, methodology and key findings. The thesis abstract expands on key contextual elements, including the limitations of intrinsically motivated methods for mitigating response biases, and references the thesis reflection section.

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 and socially desirable responses (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 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 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 honest 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, 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 unilaterally deviate from their chosen strategy.

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 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. We introduced an interim payment midway through the survey to address potential challenges such as participant scepticism and uncertainty about incentives. The interim payment served 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.

Two BTS experimental conditions were specified in investigating these aims: one without an interim payment and one with an interim payment. Participants' BTS scores for each item were calculated and summed in both conditions. As the survey was undertaken in two parts (see ‘Procedure’ section), the items were summed for each of the two parts of the survey. Both bonuses were paid at the survey’s conclusion in the former condition. 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 as the interim payment. The Regular Incentive condition served as the control group, where participants received the participation payment without 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 hypothesised that:

- H1: Participants subjected to the BTS (with or without an interim payment) would have significantly higher mean scores indicating agreement with socially undesirable statements than those in the Regular Incentive condition.
- H2: Participants subjected to the BTS with an interim payment would have significantly higher mean scores indicating agreement with socially undesirable statements than 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, n.d.-b).

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 Lakens (2022) advises. 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 α = .025, the 'α err prob' setting was specified at .05 as, by definition, an F-test is undirected. With three conditions, this sample size translates to approximately 263 participants per group. While Schoenegger (2021) estimated a 5% exclusion rate, it was anticipated that the current two-part study might 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 was launched. Once the target was met, data collection ceased without a time-based stopping rule. However, as the survey was to be completed in two parts, a time-based stopping rule was implemented for Part 2. 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.

Upon completing Part 1, participants were invited to return and complete the second part approximately 48 hours later. 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 exceeded 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 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 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, they made predictions following the main questions as in the two BTS conditions to maintain consistent base compensation per hour across 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. The study did not include attention checks, 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?" with options of 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	Question	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 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

After data cleaning, the analyses were conducted using R (R Core Team, 2024). The percentage of missing data varied slightly by condition: BTS (3.11%), BTS+IP (3.07%) and RI (2.37%) resulting in an overall missing data percentage of 2.85%. 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 generated to summarise the sample characteristics regarding 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).

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)

The contrasts were confirmed as orthogonal, as the sum of the products of weights equalled 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 analysis to quantify effect sizes for the pairwise planned contrasts.

The following inferential criteria applied:

- H1 was considered supported if the mean score was higher in the BTS condition (with or without interim payment) than in the RI condition, with $p < .025$, 1-tailed.
- H2 was considered supported if the mean score was 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 & Rouder, 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 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 were 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 over 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)

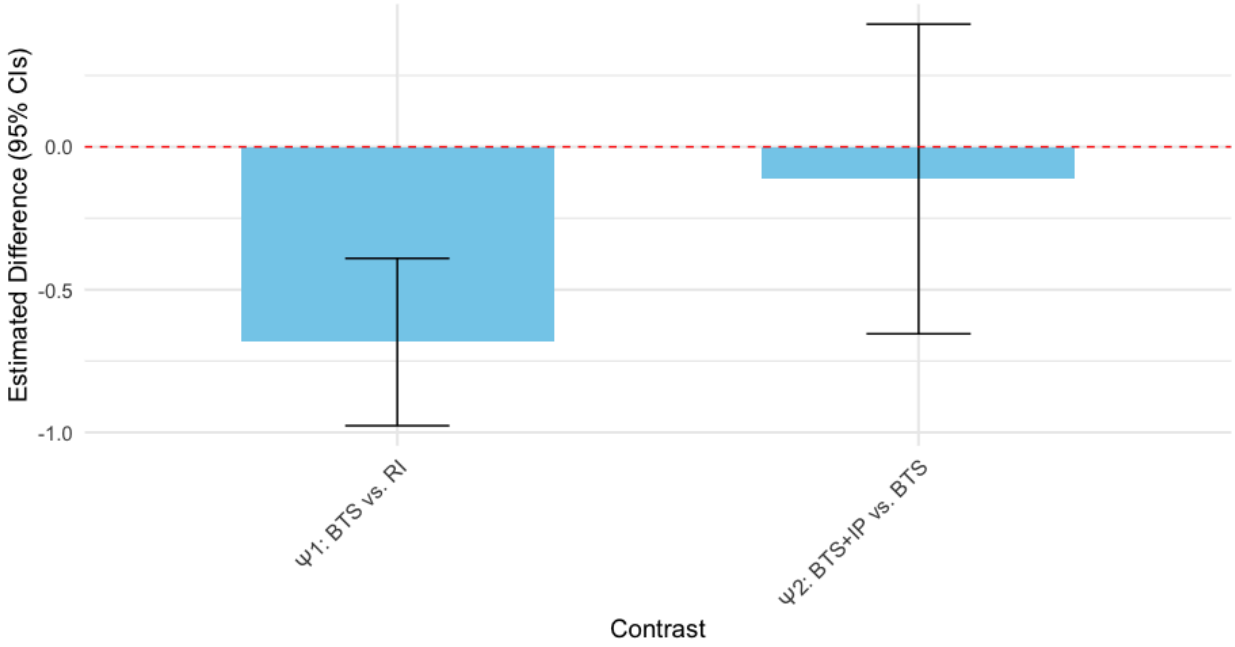
Condition	Mean	95% CI Lower	95% CI Upper	Standard Deviation
BTS	23.08	16.62	29.54	5.71
BTS+IP	22.85	16.25	29.45	5.83
RI	25.02	19.04	30.99	5.28

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. 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.

Figure 3

Mean Differences and 95% Confidence Intervals for Planned Contrasts

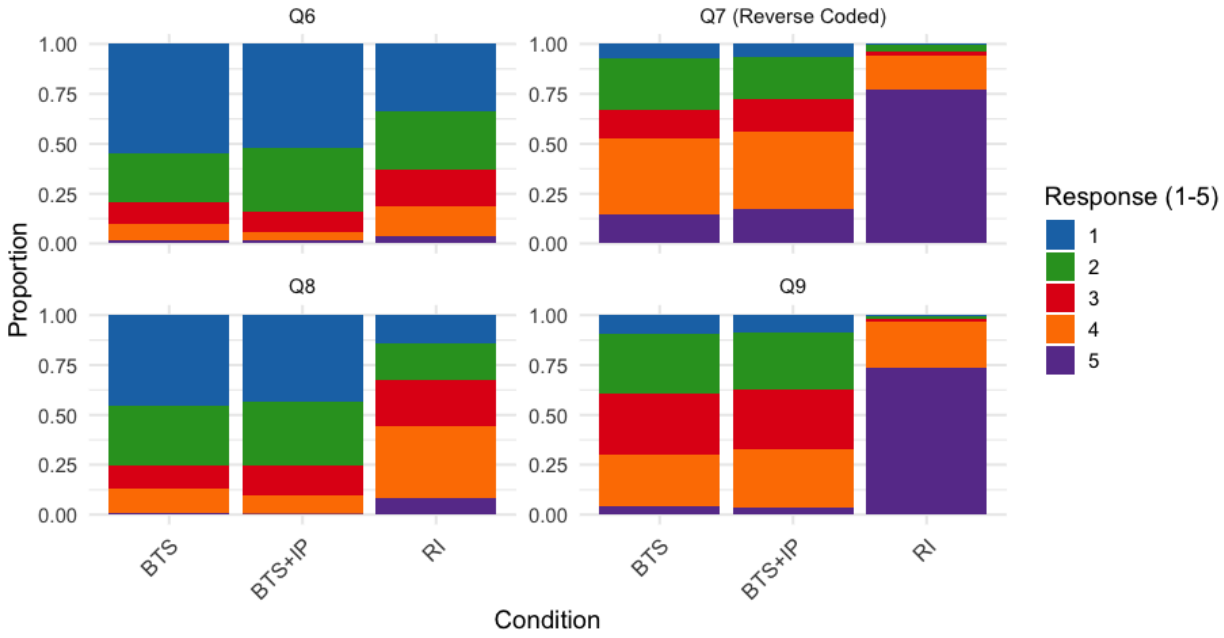


Note. Ψ1 represents the planned contrast comparing BTS (with or without interim payment) vs. Regular Incentive. Ψ2 represents the planned contrast comparing BTS with interim payment vs. BTS alone. Bars represent the estimated difference for each planned contrast, with error bars indicating 95% confidence 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.

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 ($\alpha = .005$). 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.

Figure 4



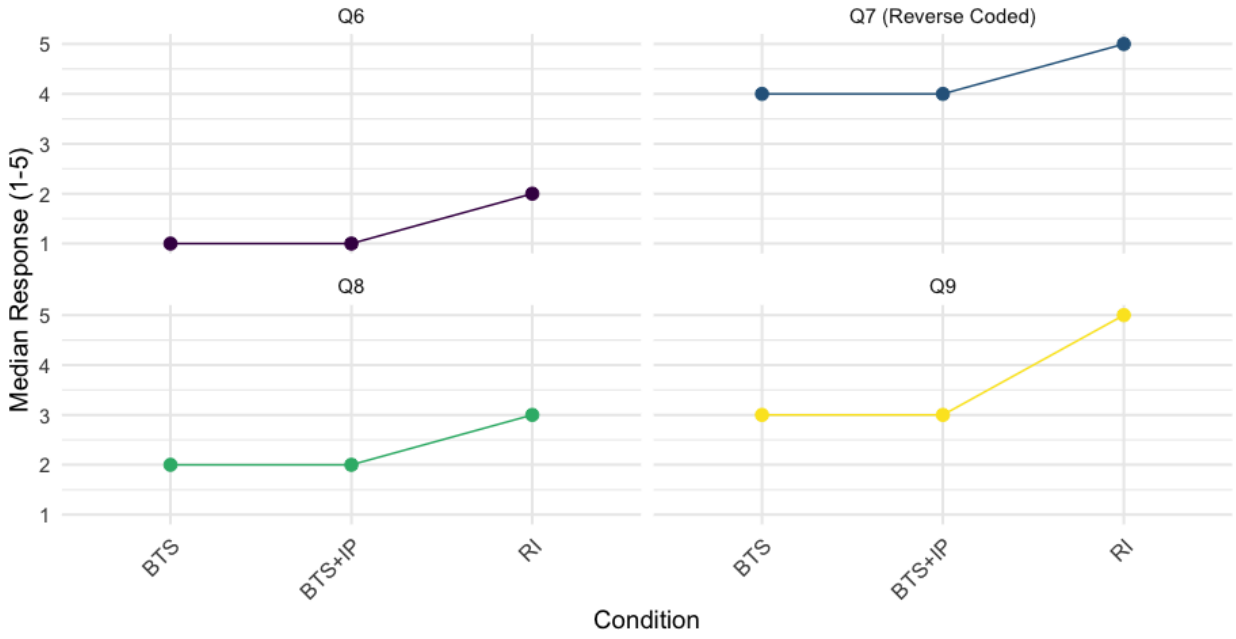
Response Distributions for Items with Significant Differences Across Conditions

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.

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 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. 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 understand better 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 additional research identifies conditions under which the BTS performs effectively, it cannot be recommended as a practical tool for applied psychological research.[3]

[3] A consolidated reference list is provided at the end of the thesis.

Reflection

A Journey of Discovery

Research is as much a process of self-discovery as it is an exploration of the subject at hand; at least, that has been my experience. In this reflection, I explore the challenges, lessons and personal growth that have shaped my thesis journey while connecting them to my broader professional aspirations of becoming a clinical psychologist. While reflection sections are uncommon in quantitative theses, it seems appropriate here, given that this study falls within meta-psychology, which examines how psychological research is conducted. Furthermore, as completing a registered report as part of a Master's thesis is rare (at least in New Zealand), this reflection may serve as a resource for students and supervisors embarking on similar projects.

I use Bayesian updating as a metaphor for my development throughout this journey. This metaphor is fitting since the focus of this research, the BTS, is grounded in Bayesian principles. Just as Bayesian updating refines beliefs based on new evidence, the registered report process reshaped my understanding of research, reinforcing key qualities: intellectual flexibility through adapting to feedback, openness to collaboration with supervisors and reviewers, evidence-based practice through pre-registration, and a commitment to lifelong learning by engaging with complex methodologies. These reflections mark both the conclusion of this thesis and the beginning of a career shaped by these values.

The Impact of the Registered Report

During the taught portion of my master's, I became increasingly aware of challenges in psychological research, particularly issues such as publication bias and

the file drawer problem (Franco et al., 2014; Rosenthal, 1979). These issues arise when studies with null results remain unpublished, often because journals prioritise positive findings. The registered report format addresses these concerns by requiring hypotheses, methods and analysis strategies to undergo peer review before data collection (Chambers, 2013; PCI, n.d.-a). Importantly, this format ensures that the study will be published regardless of the results provided the protocol is adhered to, thereby ensuring that null results are reported alongside positive ones. This approach enhances transparency and reproducibility by preventing selective reporting and clarifying that hypotheses were pre-registered rather than developed after the fact (Munafò et al., 2017). My supervisor, an advocate of registered reports, encouraged this approach, which aligned with my growing commitment to evidence-based and transparent research practices.

A key feature of the registered report format is its structured process, which presents advantages and challenges. For instance, one requirement was to present the study design plan in a concise, tabular form (PCI, n.d.-b). Additionally, with the final article limited to 5,000 words, the Stage 1 manuscript had to remain focused, with only about 3,000 words dedicated to outlining hypotheses, methods and analysis plans. These constraints encouraged me to refine my thinking, ensuring all essential information was provided with no superfluous text. Communicating complex ideas succinctly was a challenge but ultimately enhanced my ability to structure research proposals clearly and efficiently. This process not only improved the coherence of our study but also strengthened its scientific rigour by reducing ambiguity in the research design.

The registered report process also improved the study's analytical framework. Initially, I had planned to use Welch's t-tests for the primary analysis. However, one of the peer reviewers recommended planned contrasts as a more suitable alternative. This method allowed for more focused comparisons aligned with specific hypotheses, enabling a clearer test of directional predictions and greater statistical power (Field, 2018; Maxwell & Delaney, 2004). Peer feedback also shaped the inclusion of supplementary Bayesian analyses to interpret null results, which offered additional insights beyond frequentist methods. These improvements demonstrated the value of collaboration and iterative refinement in research planning, achieving a rigour that might not typically be expected in a Master's thesis. Though more demanding, these changes strengthened the study's analytical design and enhanced my capacity to adapt and improve research approaches.

Ultimately, following the registered report framework demonstrated the importance of scientific robustness and reproducibility, even in the face of null results. The findings did not support our hypotheses, as the BTS did not significantly mitigate SDR, challenging assumptions about its effectiveness. In traditional publication models, such results might have been relegated to the file drawer. However, the PCI RR framework guarantees publication for studies that adhere to their pre-registered plans, ensuring that all findings, whether confirming or challenging hypotheses, contribute to the broader scientific discourse (Nosek et al., 2015). For me, this process reinforced a deeper sense of purpose in research. Knowing that the findings would be shared regardless of their outcome affirmed that the study had value beyond personal

milestones by contributing to ongoing efforts to improve psychological research practices and understanding.

Facing and Overcoming Challenges

As mentioned earlier, the registered report journey presented various challenges that tested my ability to adapt and problem-solve as a first-time researcher. One significant challenge arose when the analysis plan expanded in response to peer reviewers' inputs. Reviewers raised questions about interpreting null results and suggested additional analyses, including item-level exploratory analyses, to better understand variability across survey responses. Addressing this feedback meant moving beyond the original plan, which focused solely on tests of mean differences. Additionally, reviewers offered differing priorities, with some advocating for Bayesian methods and others being open to either Bayesian or frequentist approaches, provided the analysis was rigorous and addressed the concerns raised. Managing these differing viewpoints was complex, particularly within the constraints of a Master's thesis. However, with my supervisor’s guidance, I learned to prioritise reviewers’ comments and suggestions strategically. For example, we deferred certain exploratory analyses to supplementary materials to maintain feasibility. This approach demonstrated respectful engagement with feedback, strengthened the study’s design and improved its chances of making a meaningful contribution to the field through publication.

Receiving peer reviews, particularly signed reviews through PCI RR, was both a challenge and a unique opportunity compared to the typical master's research experience. One of the reviewers was a researcher whose work I had cited extensively and whose registered report (Schoenegger & Verheyen, 2022) had influenced my study

design. Knowing they would evaluate my work triggered self-doubt and fear of judgment, a common experience for early-career researchers, often associated with imposter syndrome (Clance & Imes, 1978; Parkman, 2016). Unlike traditional master's projects, which may only receive feedback during final assessment, I received multiple rounds of detailed peer review early in the process. Although the commentary from four reviewers was substantial, I approached it systematically by breaking the critiques into manageable tasks and seeking my supervisor's support when needed. This iterative process taught me to embrace constructive criticism and strengthened my capacity to refine and defend my research decisions. Ultimately, the structured feedback inherent to the registered report format accelerated my development as a researcher and will continue to shape my approach to scientific collaboration and practice.

The study's unexpected results presented their own challenges, particularly in managing the doubt and uncertainty they elicited. When the findings contradicted our hypotheses, I initially questioned whether there were flaws in the preregistered method or analysis strategy. After verifying the robustness of the plan, I felt drawn to exploratory analyses in hopes of uncovering evidence that the BTS had influenced SDR. Item-level analyses revealed intriguing patterns: for one question, the BTS appeared to reduce SDR, while in others, it seemed to increase it. These findings raised questions about whether the direction of SDR might vary across cultural norms or participant groups, highlighting the complexity of defining SDR. However, the registered report framework and my supervisor’s guidance helped me refocus on the pre-registered inferential analyses to balance curiosity and methodological rigour.

This experience also provided insights into psychological concepts such as cognitive dissonance (Festinger, 1957) and motivated reasoning (Kunda, 1990). Cognitive dissonance arises when evidence contradicts beliefs, causing discomfort, much like the tension I experienced when faced with null findings that challenged my assumptions about the BTS. This discomfort initially led me to prioritise exploratory results, demonstrating motivated reasoning in action. I learned to recognise and counteract these biases by adhering to the registered report framework. Its methodical approach helped me maintain intellectual integrity, a lesson that will guide my future research and clinical practice.

Growth and Future Directions

As a mature student transitioning from a career in engineering consultancy and operations management, this thesis presented a new challenge: conducting independent scientific research. While my experience managing large-scale projects meant I was confident in my organisational skills and ability to meet deadlines, this research required a different type of discipline: one driven primarily by self-motivation rather than external pressures like organisational goals or financial imperatives. The registered report approach, particularly the scheduled review option I opted into to shorten the Stage 1 timeframe, provided momentum in the early stages of the project. The structured workflow guided my progress, keeping me engaged and on track, even during the logistical challenges of moving from Singapore to Australia. Ultimately, this approach helped me stay focused and maintain steady progress.

The registered report format also deepened my approach to learning and growth. Much like Bayesian updating, my understanding evolved as I incorporated new

information and refined my perspective. With the benefit of life experience, I deliberately stretched myself by embracing new methods and added complexities. My younger self might have approached this process with a narrower focus of simply completing the task, but I found greater value in engaging deeply with the challenges and prioritising learning over immediate results. This shift in mindset, supported by the iterative feedback and reflection inherent to the registered report approach, became one of the most rewarding aspects of this experience, reshaping how I view professional growth and intellectual development.

One of the most surprising outcomes of this journey has been the realisation that I want to remain involved in research alongside clinical practice. Initially, I viewed this thesis as a stepping stone toward becoming a clinical psychologist. However, grappling with complex methodologies, responding to peer reviews, and engaging deeply with the psychological literature have sparked a lasting interest in integrating research into my future work. Single-case experimental designs (SCEDs), for example, provide a systematic way to test interventions over time with individual clients, offering quantitative insights that enhance clinical decision-making and improve outcomes (Kazdin, 2011). This approach reflects the core principles I developed throughout this thesis: balancing curiosity with rigour, valuing feedback and applying research to real-world problems.

Beyond shaping my approach to scientific inquiry, this research has deepened my understanding of the challenges inherent in self-report methodologies and the complexities of eliciting truthful responses. By critically evaluating the BTS in a psychological context and testing a novel strategy to enhance its efficacy, this study contributes to research on bias mitigation. Engaging with the registered report format

further strengthens its contribution to ongoing efforts to improve the accuracy of self-report data in psychological research. The following conclusion consolidates these contributions, considering their broader implications for the field.

Conclusion

The study addressed a persistent issue in psychological research: the impact of response biases, particularly SDR, on the validity of self-reports in sensitive contexts. While self-report measures are essential for capturing internal states and personal experiences, they are vulnerable to bias. Researchers have explored various mitigation strategies, including incentive-compatible mechanisms that align participants' motivations with accurate reporting. The BTS is one such method, theoretically incentivising honesty through its scoring system, which rewards surprisingly common responses. However, despite its conceptual appeal, practical concerns, particularly regarding trust in the mechanism and its limited application in psychology, raised questions about its effectiveness.

This study empirically tested the BTS in a psychological research setting, building on prior evaluations (e.g., Schoenegger & Verheyen, 2022). Rather than increasing agreement with socially undesirable statements, the BTS conditions resulted in lower agreement compared to the regular incentive condition, reflecting increased rather than reduced SDR. Adding an interim payment, designed to enhance trust in the BTS mechanism, did not significantly alter these findings. As the second registered report to critically assess the BTS, this study challenges its reliability as a tool for mitigating SDR in self-report data.

These findings highlight the ongoing challenge of response biases and the need for continued innovation in methods for eliciting truthful responses. Until a robust, empirically validated solution emerges, self-report biases will remain a significant concern. In mental health research, individuals often underreport depressive symptoms due to societal expectations, complicating assessment and treatment decisions (Latkin et al., 2017). Similarly, studies on substance use show that individuals understate alcohol consumption due to social judgment, leading to misleading prevalence estimates that can impact public health policies (Davis et al., 2010). Beyond behavioural self-reports, SDR also distorts attitudinal measures. For example, research on conspiracy theory endorsement suggests that some individuals report belief in conspiracy theories without genuine conviction, skewing findings in this area (Ross et al., 2024). These examples reinforce the need for improved strategies to reduce response biases and ensure that self-report data better inform clinical practice, research, policymaking and public discourse.

While this study does not offer a definitive solution, it contributes to a broader understanding of the complexities of incentivising truthful responding. Ensuring that studies like this are accessible to the scientific community is essential for meaningful progress. The registered report format plays a crucial role in advancing transparent research practices by reducing publication bias and strengthening study design through pre-registration and peer review. By reflecting on the challenges and benefits of this format, this thesis may serve as a resource for students and supervisors considering registered reports for master's-level research.

Ultimately, this study updates the field's priors on the efficacy of the BTS and the broader challenge of response biases in self-report data. It reinforces the need for methodological advancements to enhance data reliability in psychology. As psychological research continues to rely on self-report measures, the demand for robust, evidence-based solutions to response biases remains critical. By critically evaluating the BTS within a registered report framework, this study contributes to the ongoing effort to improve the validity of self-report measures and highlights the necessity of transparent, reproducible research. While the BTS may not provide the solution once hoped for, this research represents a step toward refining methodologies that enhance the credibility of psychological data. The challenge of eliciting truthful responses is far from resolved, but each inquiry, whether yielding supportive or null findings, adds to the collective effort to advance psychological science.

References

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

Alvarez, R. M., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying attention to inattentive survey respondents. *Political Analysis*, *27*(2), 145–162. https://doi.org/10.1017/pan.2018.57

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, *10*(3), 229–240. https://doi.org/10.1037/1089-2680.10.3.229

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.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*(6), 1014–1027. https://doi.org/10.1037/0022-3514.81.6.1014

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

Becker, T. E., & Colquitt, A. L. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. *Personnel Psychology*, *45*(2), 389–406. https://doi.org/10.1111/j.1744-6570.1992.tb00855.x

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

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2013). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*, 23–36.

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.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229. https://doi.org/10.1037/pspp0000085

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

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

Campbell, D. T. (1950). The indirect assessment of social attitudes. *Psychological Bulletin*, *47*(1), 15–38. https://doi.org/10.1037/h0054114

Carson, R. T. (2012). Contingent valuation: A practical alternative when prices aren't available. *Journal of Economic Perspectives*, *26*(4), 27–42. https://doi.org/10.1257/jep.26.4.27

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology*, *60*(6), 767–769. https://doi.org/10.1037/0021-9010.60.6.767

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. Lance & R. Vandenberg (Eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 309–336). Routledge/Taylor & Francis Group.

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

Cialdini, R. B. (1984). *Influence*. William Morrow.

Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice*, *15*(3), 241–247. https://doi.org/10.1037/h0086006

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

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. Wiley.

Cummings, R. G., & Harrison, G. W. (1995). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive-compatible? *American Economic Review*, *85*(1), 260–266.

Cummings, R. G., & Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, *89*(3), 649–665. https://doi.org/10.1257/aer.89.3.649

Dalal, D. K., & Hakel, M. D. (2016). Experimental comparisons of methods for reducing deliberate distortions to self-report measures of sensitive constructs. *Organizational Research Methods*, *19*(3), 475–505. https://doi.org/10.1177/1094428116639131

Davis, C. G., Thake, J., & Vilhena, N. (2010). Social desirability biases in self-reported alcohol consumption and harms. *Addictive Behaviors*, *35*(4), 302–311. https://doi.org/10.1016/j.addbeh.2009.11.001

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

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer.

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

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2017). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, *60*(3), 340–360. https://doi.org/10.1177/00131640021970583

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Wadsworth Cengage Learning.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118. https://www.aeaweb.org/articles?id=10.1257/jep.10.3.103

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

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

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303–315. https://doi.org/10.1086/209351

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, *17*(2), 105–120. https://doi.org/10.1002/(sici)1520-6793(200002)17:2%3C105::aid-mar3%3E3.0.co;2-9

Fisher, R., & Teliis, G. (1998). Removing social desirability bias with indirect questioning: Is the cure worse than the disease? *Advances in Consumer Research*, *25*.

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

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall/Crc.

Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*, *24*(6), 746–762. https://doi.org/10.1177/1073191115624547

Hellhammer, D. H., Wüst, S., & Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*, *34*(2), 163–171. https://doi.org/10.1016/j.psyneuen.2008.10.026

Ho, C.-M., Slivkins, A., Suri, S., & Vaughan, J. (2015). *Incentivizing high quality crowdwork*. WWW 2015, Florence, Italy. https://doi.org/10.1145/2736277.2741102

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

Howie, P. J., Wang, Y., & Tsai, J. (2010). Predicting new product adoption using Bayesian truth serum. *Journal of Medical Marketing*. https://doi.org/10.1057/jmm.2010.19

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2011). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Jacquemet, N., Joule, R. V., Luchini, S., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, *65*(1), 110–132. https://doi.org/10.1016/j.jeem.2012.05.004

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

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

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 433–438. https://doi.org/10.3758/bf03200723

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

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

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Ladenburg, J., & Olsen, S. B. (2014). Augmenting short cheap talk scripts with a repeated opt-out reminder in choice experiment surveys. *Resource and Energy Economics*, *37*, 39–63. https://doi.org/10.1016/j.reseneeco.2014.05.002

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

Latkin, C. A., Edwards, C., Davey-Rothwell, M. A., & Tobin, K. E. (2017). The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. *Addictive Behaviors*, *73*(1), 133–136. https://doi.org/10.1016/j.addbeh.2017.05.005

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

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299. https://doi.org/10.1016/j.jesp.2012.07.002

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

Mamkhezri, J., Thacher, J. A., Chermak, J. M., & Berrens, R. P. (2020). Does the solemn oath lower WTP responses in a discrete choice experiment application to solar energy? *Journal of Environmental Economics and Policy*, *9*(4), 447–473. https://doi.org/10.1080/21606544.2020.1738276

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376. https://doi.org/10.1111/j.1468-0262.2004.00537.x

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data*. Routledge.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

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

Mensch, B. S., & Kandel, D. B. (1988). Underreporting of substance use in a national longitudinal youth cohort: Individual and interviewer effects. *Public Opinion Quarterly*, *52*(1), 100. https://doi.org/10.1086/269084

Mick, D. G. (1996). Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism. *Journal of Consumer Research*, *23*(2), 106. https://doi.org/10.1086/209470

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, *51*(9), 1359–1373. https://doi.org/10.1287/mnsc.1050.0379

Miller, R., Plessow, F., Kirschbaum, C., & Stalder, T. (2013). Classification criteria for distinguishing cortisol responders from nonresponders to psychosocial stress. *Psychosomatic Medicine*, *75*(9), 832–840. https://doi.org/10.1097/psy.0000000000000002

Mitchell, R. C., & Carson, R. T. (1989). *Using surveys to value public goods the contingent valuation method*. Resources for the Future.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

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

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). https://doi.org/10.1038/s41562-016-0021

Nash, J. F. (1950). The bargaining problem. *Econometrica*, *18*(2), 155–162. https://doi.org/10.2307/1907266

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

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

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., & Briggs, J. C. (2018). Registered replication report: Dijksterhuis and van knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268–294. https://doi.org/10.1177/1745691618755704

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist*, *56*(9), 717–734. https://doi.org/10.1037/0003-066x.56.9.717

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

Parkman, A. (2016). The imposter phenomenon in higher education: Incidence and impact. *Journal of Higher Education Theory and Practice*, *16*(1). https://articlegateway.com/index.php/JHETP/article/view/1936/1836

Pashler, H., Rohrer, D., & Harris, C. R. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology*, *49*(6), 959–964. https://doi.org/10.1016/j.jesp.2013.05.011

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.

Paulhus, D., & Vazire, S. (2007). The self-report method. In R. Robins, R. Fraley, & R. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 224–239). Guilford Press.

Peer Community In [PCI]. (n.d.-a). *About*. Rr.peercommunityin.org. Retrieved February 11, 2025, from https://rr.peercommunityin.org/about/about

Peer Community In [PCI]. (n.d.-b). *Guide for authors*. Rr.peercommunityin.org. Retrieved February 11, 2025, from https://rr.peercommunityin.org/help/guide_for_authors#h_2751396573533161330 9625021

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*. https://doi.org/10.3758/s13428-021-01694-3

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

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

Prolific. (2024a). *Easily find vetted research participants and AI taskers at scale*. https://www.prolific.com

Prolific. (2024b, March). *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

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578. https://doi.org/10.1016/j.jesp.2006.05.011

Qualtrics. (2024). *Transform insight into impact*. 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/

Rasinski, K. A., Visser, P. S., Zagatsky, M., & Rickett, E. M. (2005). Using implicit goal priming to improve the quality of self-report data. *Journal of Experimental Social Psychology*, *41*(3), 321–327. https://doi.org/10.1016/j.jesp.2004.07.001

Rosenberg, M. (1965). Rosenberg self-esteem scale. *PsycTESTS Dataset*, *1*(1). https://doi.org/10.1037/t01038-000

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* Appleton-Century-Crofts.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

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

Ross, R., Gleeson, K., Wilson, S., Ashton, L., & Levy, N. (2024). *Do people sincerely believe conspiracy theories that they endorse?* In PsyArXiv Preprints. 10.31234/osf.io/zsncr

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/AMR.1998.926617

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*. https://doi.org/10.1007/s11229-021-03276-4

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

Schönegger, 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

Simunović, V., & žeželj, I. (2023). Managing self-presentation. *Studia Psychologica*, *65*(2), 103–119. https://doi.org/10.31577/sp.2023.02.869

Singer, E., & Ye, C. (2012). 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. F. (1953). *Science and human behavior*. Macmillan.

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

Toulis, P., Parkes, D. C., Pfeffer, E., & Zou, J. (2015). *Incentive-Compatible experimental design*. https://doi.org/10.1145/2764468.2764525

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. (2011). *Belief elicitation: A horse race among truth serums*. CORE.

Trautmann, S. T., & van de Kuilen, G. (2015). 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

van de Morte, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, *25*(4), 40–44.

van de Schoot, R., Winter, S. D., Griffioen, E., Grimmelikhuijsen, S., Arts, I., Veen, D., Grandfield, E. M., & Tummers, L. G. (2021). The use of questionable research practices to survive in academia examined with expert elicitation, prior-data conflicts, Bayes factors for replication effects, and the Bayes truth serum. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.621547

Vecchio, R., Caso, G., Cembalo, L., & Borrello, M. (2020). Is respondents' inattention in online surveys a major issue for research? *Economia Agro-Alimentare*, *1*, 1–18. https://doi.org/10.3280/ecag1-2020oa10069

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

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

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

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), 1–10. https://doi.org/10.1038/s41598-024-51653-z

Witkowski, J. (2014). *Robust peer prediction mechanisms* [Dissertation].

Yang, M., Jia, C., & Wang, Z. (2016). Performance comparison of two truth telling incentive mechanisms: An experimental method. *2016 IEEE IEEM*. https://doi.org/10.1109/ieem.2016.7798045

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

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *The Academy of Management Review*, *12*(2), 250. https://doi.org/10.2307/258533

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. https://doi.org/10.1016/j.eneco.2019.104484

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

Zizzo, D. J. (2009). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. https://doi.org/10.1007/s10683-009-9230-z

Appendix A

Survey Questionnaire

Qualtrics Survey Software
30/10/24, 11:21 AM

Information Sheet

Welcome to our study!

My name is Claire Neville, and I'm conducting this research as part of my Master's Degree in Science and Technology (Psychology) at Massey University of New Zealand. I'm excited to have you consider participating in this important study.

Project Description and Invitation

In this study, we're investigating the Bayesian Truth Serum (BTS), a survey scoring method designed to improve the accuracy of self-report data. The BTS is a tool that can enhance the reliability of psychology research outcomes by encouraging truthful responses, particularly in situations where the truth cannot be objectively verified. Your participation is invaluable as it can significantly contribute to our understanding of how the BTS method works and its potential benefits for psychology research. You don't need an in-depth knowledge of the BTS mechanism to participate. However, if you're interested, you can read this paper by Prelec (2004), who first proposed the BTS.

Participant Identification and Recruitment

Participants are being recruited through Prolific (prolific.com), where the minimum age requirement is automatically set at 18 years and above. Participants from the US, Canada, UK, Ireland, Australia, and NZ are eligible to participate. Selection criteria include fluent English proficiency and the completion of at least 20 previous surveys on Prolific. Your name will not be

collected for this research. However, please be aware that your Prolific ID may be associated with your survey responses to verify bonus eligibility and process payments.

What to Expect

Survey Overview

You'll be asked to respond to 10 main questions covering a range of topics, including social dominance, gender-based inequalities and ageism. In addition to the main questions, we'll also collect demographic information through three short questions. The survey will be completed in two parts. You will be invited to return and complete the second part approximately 48 hours after the first part. You will then have 72 hours to complete the survey.

Potential Discomfort

Because these questions touch on sensitive social issues, you may briefly experience discomfort. However, considering the nature of the questions and the short duration of the survey, any psychological discomfort is expected to be temporary and mild.

Support Resources

If you feel distressed during or after participating, please don't hesitate to reach out to specialist resources:
Crisis Text Line: Available in the US, Canada, UK, and Ireland. https://www.crisistextline.org/
In the US, text HOME to 741741
In Canada, text CONNECT to 686868
In the UK, text SHOUT 85258
In Ireland, text HELLO to 50808
Lifeline: Available in Australia and New Zealand.
Australia: https://www.lifeline.org.au/

New Zealand: https://www.lifeline.org.nz/

Compensation

You'll receive a baseline payment of £1 for completing the full survey, with the amount split equally between Parts 1 and 2. Additionally, you may have the chance to earn a small bonus payment. Details will be provided within the survey.

Data Management

As this research is being undertaken in accordance with the principles of open science, anonymised data will be openly shared. Initially, the raw data will be password-protected and accessible only to the research team. This raw data will not contain any directly identifying information, although it will contain your Prolific ID and IP address. Once we have analysed the data to answer our research questions, we will remove your Prolific ID and your IP address before sharing the data openly on the Open Science Framework (where it will be stored indefinitely). This will allow other researchers and members of the public to access the data for validation and potentially explore new research questions.

Project Contacts

If you have any questions about the project, feel free to contact us via Prolific or by email via the contact information below:

Researcher: Claire Neville,
claire.neville.1@uni.massey.ac.nz
Supervisor: Dr. Matt Williams,
m.n.williams@massey.ac.nz

Massey University Human Ethics Committee Approval Statement

This project has been reviewed and approved by the Massey University Human Ethics Ohu Matatika 2, Application OM2 24/26. If you have any concerns about the conduct of this research, please contact the Chairperson, Massey University Human Ethics Ohu Matatika 2, email humanethics2@massey.ac.nz.

Thank you for considering participation in our study!

Consent Item

Having read and understood the information provided above, do you consent to participate in this study?

Yes
No

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

${e://Field/PROLIFIC_PID}

Demographic Questions

How old are you?

Under 18
18-24 years old
25-34 years old
35-44 years old
45-54 years old
55-64 years old
65+ years old

How do you describe yourself?
Male
Female
Non-binary / third gender
Prefer to self-describe
Prefer not to say

What is the highest level of education you have completed?
Less than Primary
Primary
Some Secondary
Secondary
Vocational or Similar
Some University but no degree
University - Bachelors Degree
Graduate or professional degree (MA, MS, MBA, PhD, Law Degree, Medical Degree etc)
Prefer not to say

BTS Text - BTS Condition

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.

I have read the above message and understand that the top 50% of participants, ranked by their truthfulness scores, will be eligible for a bonus.
Yes
No

BTS_Q1
Some groups of people are simply inferior to other groups.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q2
An ideal society requires some groups to be on top and others to be on the bottom.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your

estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q3
Group equality should not be our primary goal.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q4
It is unjust to try to make groups equal.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q5

All in all, men have more responsibilities and fewer benefits.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS Text - BTS + IP Condition

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.

I have read the above message and understand that the top 50% of participants, ranked by their truthfulness scores, will be eligible for a bonus.
Yes
No

BTS_IP_Q1
Some groups of people are simply inferior to other groups.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q2
An ideal society requires some groups to be on top and others to be on the bottom.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q3

Group equality should not be our primary goal.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q4

It is unjust to try to make groups equal.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q5
All in all, men have more responsibilities and fewer benefits.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI Condition
Pre-Survey Text
You will now proceed to the survey questions. Please answer all questions truthfully.

RI_Q1

Some groups of people are simply inferior to other groups.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q2

An ideal society requires some groups to be on top and others to be on the bottom.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q3

Group equality should not be our primary goal.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q4

It is unjust to try to make groups equal.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q5

All in all, men have more responsibilities and fewer benefits.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

End of Survey Message

Thank you for taking part in our survey. Please click the button below to be redirected back to Prolific and register your submission.

Information Sheet

Welcome to Part 2 of our study.

Thank you for returning to complete Part 2 of our study, which is investigating the Bayesian Truth Serum (BTS), a survey scoring method designed to improve the accuracy of self-report data.

Remember, if you feel distressed during or after participating, please don't hesitate to reach out to specialist resources:

Crisis Text Line: Available in the US, Canada, UK, and Ireland.

https://www.crisistextline.org/

In the US, text HOME to 741741
In Canada, text CONNECT to 686868
In the UK, text SHOUT 85258
In Ireland, text HELLO to 50808

Lifeline: Available in Australia and New Zealand.

Australia: https://www.lifeline.org.au/
New Zealand: https://www.lifeline.org.nz/

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

${e://Field/PROLIFIC_PID}

BTS Text - Part 2

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.

I have read the above message and understand that the top 50% of participants, ranked by their truthfulness scores, will be eligible for a bonus.

Yes
No

Quality Check

Based on the instructions provided, what percentage of participants, ranked by their truthfulness scores, will be eligible for a bonus?

30%
50%
100%

BTS_Q6

Nowadays, men don’t have the same chances in the job market as women.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q7

Men are not particularly discriminated against.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q8

Doctors spend too much time treating sickly, older people.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q9

Younger people are usually more productive than older people at their jobs

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_Q10

Older people don’t really need to get the best seats on buses and trains.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

End of Survey Message

Thank you for taking part in our survey. Please click the button below to be redirected back to Prolific and register your submission.

Information Sheet

Welcome to Part 2 of our study.

Thank you for returning to complete Part 2 of our study, which is investigating the Bayesian Truth Serum (BTS), a survey scoring method designed to improve the accuracy of self-report data.

Remember, if you feel distressed during or after participating, please don't hesitate to reach out to specialist resources:

Crisis Text Line: Available in the US, Canada, UK, and Ireland.

https://www.crisistextline.org/

In the US, text HOME to 741741
In Canada, text CONNECT to 686868
In the UK, text SHOUT 85258
In Ireland, text HELLO to 50808

Lifeline: Available in Australia and New Zealand.

Australia: https://www.lifeline.org.au/
New Zealand: https://www.lifeline.org.nz/

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

${e://Field/PROLIFIC_PID}

BTS Text - BTS + IP Condition Part 2

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.

I have read the above message and understand that the top 50% of participants, ranked by their truthfulness scores, will be eligible for a bonus.

Yes
No

Quality Check

What percentage of participants, ranked by their truthfulness scores, will be eligible for a bonus?

30%
50%
100%

BTS_IP_Q6

Nowadays, men don’t have the same chances in the job market as women.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q7

Men are not particularly discriminated against.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q8

Doctors spend too much time treating sickly, older people.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q9

Younger people are usually more productive than older people at their jobs

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

BTS_IP_Q10

Older people don’t really need to get the best seats on buses and trains.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

End of Survey Message

Thank you for taking part in our survey. Please click the button below to be redirected back to Prolific and register your submission.

Information Sheet

Welcome to Part 2 of our study.

Thank you for returning to complete Part 2 of our study, which is investigating the Bayesian Truth Serum (BTS), a survey scoring method designed to improve the accuracy of self-report data.

Remember, if you feel distressed during or after participating, please don't hesitate to reach out to specialist resources:

Crisis Text Line: Available in the US, Canada, UK, and Ireland.

https://www.crisistextline.org/

In the US, text HOME to 741741
In Canada, text CONNECT to 686868
In the UK, text SHOUT 85258
In Ireland, text HELLO to 50808

Lifeline: Available in Australia and New Zealand.

Australia: https://www.lifeline.org.au/
New Zealand: https://www.lifeline.org.nz/

Prolific ID

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID

${e://Field/PROLIFIC_PID}

Pre-survey Text
You will now proceed to the survey questions. Please answer all questions truthfully.

RI_Q6
Nowadays, men don’t have the same chances in the job market as women.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q7
Men are not particularly discriminated against.
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.
Strongly disagree 0
Somewhat disagree 0

Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q8

Doctors spend too much time treating sickly, older people.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q9

Younger people are usually more productive than older people at their jobs.

Strongly disagree
Somewhat disagree
nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

RI_Q10

Older people don’t really need to get the best seats on buses and trains.

Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree

Out of the following options, what percentage of other participants in this study do you think chose this option as the answer to the above question? Remember, your estimates have to sum up to 100%.

Strongly disagree 0
Somewhat disagree 0
Neither agree nor disagree 0
Somewhat agree 0
Strongly agree 0
Total 0

End of Survey Message

Thank you for taking part in our survey. Please click the button below to be redirected back to Prolific and register your submission.

Appendix B

Supplementary Materials - Post hoc Analyses

Background and Methodology

This supplementary materials document reports additional analyses that are not included in the main text for brevity. The exploratory analyses included both pre-registered and post hoc components to provide a comprehensive understanding of the data:

- Pre-registered Analysis (Chi-Square Tests of Independence): The pre-registered Chi-Square tests of independence were performed using the chisq.test() in R to examine differences in response distributions across conditions for the ten survey items. These tests aimed to explore item-level variability following null results in the primary analysis. A Bonferroni correction ($\alpha = .005$) was applied to control for multiple comparisons.
- Post-hoc Analyses:
    - Permutation Tests for Central Tendency: These tests were undertaken because they are powerful and flexible in detecting shifts in central tendency (mean, median or mode) across conditions. Using the oneway_test function with the Fisher-Pitman permutation test and 9,999 resamples, this test expanded on the Chi-Square analysis to investigate significant item-level variability.
    - Brown-Mood Median Tests: These tests focused on median differences, providing a simpler and more interpretable comparison of central tendencies. Using the median_test function with 9,999 resamples, they complemented the permutation tests and validated the significant differences observed.

- Descriptive Analysis of Medians: Median responses for items Q6, Q7, Q8 and Q9 identified as significant in earlier tests were descriptively analysed. Medians were calculated for each condition and visualised to clearly and intuitively compare central tendencies across groups.
- Response Duration Analysis: Welch's t-tests were undertaken to compare response durations between conditions (BTS and BTS+IP separately), the RI condition for Part 1 and Part 2, and the total survey completion time. This analysis was prompted by unexpected findings for Q6, Q8 and Q9, where higher socially desirable responding (SDR) in the BTS conditions indicated a possible backfire effect.

Results

Chi-Square Tests of Independence

Chi-square tests of independence were performed to examine differences in response distributions across conditions for all ten survey items. Results showed significant differences for four items (Q6, Q7, Q8 and Q9), with non-significant effects for the remaining six (Q1–Q5 and Q10).

- Key Findings:
    - Q6, Q8, Q9: Effects were observed in the opposite direction to the primary hypotheses, with higher SDR in the BTS conditions.
    - Q7: The response distribution aligned with the intended effect of the BTS, showing lower SDR in the BTS conditions.
    - Q1–Q5, Q10: No significant effects were observed.
- Supporting Tables and Figures:
    - Table S1: Full Chi-Square test results, including test statistics and p-values.
    - Tables S2–S11: Contingency tables displaying response distributions by condition for each survey item.
    - Figure S2: Plot showing the proportional distribution of responses across conditions.

Table B1

Chi-Square Test Results for Response Distributions

Item	Item Statement	χ^2	*p*-value
Q1	Some groups of people are simply inferior to other groups.	10.921	.206
Q2	An ideal society requires some groups to be on top and others to be on the bottom.	17.325	.027
Q3	Group equality should not be our primary goal.	5.740	.676
Q4	It is unjust to try to make groups equal.	7.418	.492
Q5	All in all, men have more responsibilities and fewer benefits.	11.099	.196
Q6	Nowadays, men don't have the same chances in the job market as women.	50.664*	< .001
Q7	Men are not particularly discriminated against. (R)	309.438*	< .001
Q8	Doctors spend too much time treating sickly older people.	168.853*	< .001
Q9	Younger people are usually more productive than older people at their jobs.	499.653*	< .001
Q10	Older people don't really need to get the best seats on buses and trains.	16.258	.039

Note: (R) indicates reverse-coded item. χ^2 = Chi-Square statistic. * Significant results were determined using a Bonferroni-corrected threshold of $\alpha = .005$.

Table B2

Response Distribution for Q1: Some groups of people are simply inferior to other groups.

Response Condition	1	2	3	4	5
BTS	124	50	23	64	10
BTS+IP	125	41	26	71	12
RI	131	58	35	47	10

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 10.921$, $p = .206$.

Table B3

Response Distribution for Q2: An ideal society requires some groups to be on top and others to be on the bottom.

Response Condition	1	2	3	4	5
BTS	58	94	30	74	15
BTS+IP	61	78	41	79	16
RI	82	60	48	79	12

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 17.325$, $p = .027$.

Table B4

Response Distribution for Q3: Group equality should not be our primary goal.

Response Condition	1	2	3	4	5
BTS	52	76	46	64	33
BTS+IP	47	89	53	59	27
RI	51	89	42	74	25

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 5.740$, $p = .676$.

Table B5

Response Distribution for Q4: It is unjust to try to make groups equal.

Response Condition	1	2	3	4	5
BTS	94	82	32	50	13
BTS+IP	105	83	39	39	9
RI	85	92	36	57	11

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 7.418$, $p = .492$

Table B6

Response Distribution for Q5: All in all, men have more responsibilities and fewer benefits.

Response Condition	1	2	3	4	5
BTS	113	75	43	29	11
BTS+IP	106	95	35	27	12
RI	138	72	41	22	8

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 11.099$, $p = .196$.

Table B7

Response Distribution for Q6: Nowadays, men don't have the same chances in the job market as women.

Response Condition	1	2	3	4	5
BTS	148	67	29	22	5
BTS+IP	143	89	28	11	4
RI	94	82	52	42	11

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 50.664$, $p < .001$.

Table B8

Response Distribution for Q7: Men are not particularly discriminated against. (R)

Response Condition	1	2	3	4	5
BTS	20	69	40	103	39
BTS+IP	18	57	45	107	48
RI	1	9	6	49	216

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 309.438$, $p < .001$.

Table B9

Response Distribution for Q8: Doctors spend too much time treating sickly older people.

Response Condition	1	2	3	4	5
BTS	124	80	31	33	3
BTS+IP	120	88	41	25	1
RI	39	53	64	102	23

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 168.853$, $p < .001$.

Table B10

Response Distribution for Q9: Younger people are usually more productive than older people at their jobs.

Response Condition	1	2	3	4	5
BTS	25	82	83	69	12
BTS+IP	24	78	82	82	9
RI	1	4	5	64	207

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 499.653$, $p < .001$.

Table B11

Response Distribution for Q10: Older people don't really need to get the best seats on buses and trains.

Response Condition	1	2	3	4	5
BTS	118	111	19	19	4
BTS+IP	113	109	30	23	0
RI	90	134	35	19	3

Note. Responses refer to a 5-point Likert scale where 1 = "Strongly Disagree" and 5 = "Strongly Agree." $\chi^2(8) = 16.258$, $p = .039$.

Permutation Tests for Central Tendency

Permutation tests were undertaken to assess shifts in central tendency (mean, median or mode) across conditions for all ten survey items. These tests produce a Chi-Square statistic because the permutation procedure calculates test statistics for each resampling iteration. The observed statistic is then compared to the null distribution generated through permutations. While the test does not rely on theoretical distributions, the resulting statistic is expressed as a Chi-Square value to facilitate interpretation and comparison.

The results revealed significant differences for four items (Q6, Q7, Q8 and Q9), consistent with the findings from the Chi-Square analysis. No significant effects were observed for the remaining six items (Q1–Q5 and Q10).

- Key Findings:
    - Q6, Q8, Q9: Significant shifts in central tendency were observed, consistent with effects in the opposite direction to the primary hypotheses, showing higher SDR in the BTS conditions.
    - Q7: Results aligned with the intended effect of the BTS, showing lower SDR in the BTS conditions.
    - Q1–Q5, Q10: No significant shifts in central tendency were observed.
- Supporting Tables:
    - Table S12: Full permutation test results, including test statistics and p-values

Table B12

Permutation Test Results for Central Tendency

Item	Item Statement	χ^2	*p*-value
Q1	Some groups of people are simply inferior to other groups.	2.8254	.247
Q2	An ideal society requires some groups to be on top and others to be on the bottom.	1.0163	.603
Q3	Group equality should not be our primary goal.	0.4497	.806
Q4	It is unjust to try to make groups equal.	4.2168	.124
Q5	All in all, men have more responsibilities and fewer benefits.	4.4802	.104
Q6	Nowadays, men don't have the same chances in the job market as women.	44.247*	< .001
Q7	Men are not particularly discriminated against. (R)	221.87*	< .001
Q8	Doctors spend too much time treating sickly older people.	163.09*	< .001
Q9	Younger people are usually more productive than older people at their jobs.	384.36*	< .001
Q10	Older people don't really need to get the best seats on buses and trains.	3.9825	.139

Note: (R) indicates reverse-coded item. χ^2 = Chi-Square statistic. * Significant results were determined using a Bonferroni-corrected threshold of α = .005.

Brown-Mood Median Tests

Brown-Mood median tests examined median differences across conditions for all ten survey items. Results showed significant median differences for four items (Q6, Q7, Q8 and Q9), aligning with findings from the Chi-Square and permutation tests. The remaining six items (Q1–Q5 and Q10) showed no significant differences.

- Key Findings:
    - Q6, Q7, Q8, Q9: Significant median differences were observed, consistent with previous analyses.
    - Q1–Q5, Q10: No significant median differences were detected.
- Supporting Tables:
    - Table S13: Full Brown-Mood test results, including test statistics and p-values.

Table B13

Brown-Mood Median Test Results

Item	Item Statement	χ^2	*p*-value
Q1	Some groups of people are simply inferior to other groups.	2.8742	.239
Q2	An ideal society requires some groups to be on top and others to be on the bottom.	2.2463	.601
Q3	Group equality should not be our primary goal.	1.4823	.478
Q4	It is unjust to try to make groups equal.	1.8120	.404
Q5	All in all, men have more responsibilities and fewer benefits.	2.0726	.357
Q6	Nowadays, men don't have the same chances in the job market as women.	38.746*	< .001
Q7	Men are not particularly discriminated against. (R)	296.92*	< .001
Q8	Doctors spend too much time treating sickly older people.	142.06*	< .001
Q9	Younger people are usually more productive than older people at their jobs.	452.43*	< .001
Q10	Older people don't really need to get the best seats on buses and trains.	2.3231	.318

Note: (R) indicates reverse-coded item. χ^2 = Chi-Square statistic. * Significant results were determined using a Bonferroni-corrected threshold of α = .005.

Descriptive Analysis of Medians

Medians for Q6, Q7, Q8 and Q9 were calculated and visually compared across conditions to aid interpretation.

- Key Findings:
    - Q6, Q8, Q9: Medians were lower in the BTS and BTS+IP conditions than RI.
    - Q7: Median responses were higher in the BTS and BTS+IP conditions compared to RI. However, this item was reverse coded, indicating higher SDR in the BTS and BTS+IP conditions.
- Supporting Tables and Figures:
    - Median values for each condition are presented in Table S14 and graphically in Figure S2.

Table B14

Median Values for Survey Questions 6, 7, 8 and 9

Item	BTS	BTS+IP	RI
Q6	1	1	2
Q7 (R)	4	4	5
Q8	2	2	3
Q9	3	3	5

Note. (R) = Reverse coded item.

Response Duration Analysis

Longer response times were examined as an indicator of increased cognitive engagement in the BTS conditions. The duration results suggest a potential increase in cognitive engagement in BTS during Part 1, but there is no consistent pattern for Part 2 or total completion times.

- Key Findings:
    - Part 1 (BTS vs RI): BTS participants spent significantly more time on the survey than RI participants.
    - Part 1 (BTS+IP vs RI): No significant differences
    - Part 2: No significant differences across any conditions.
    - Total Duration: No significant differences across any conditions.
- Supporting Tables: Full results of Welch's t-tests are presented in Table S14.

Table B15

Results of Welch's T-tests for Response Duration Across Survey Parts and Conditions

Comparison	Part	Mean Difference (s)	Confidence Interval (s)	*t*-value	*p*-value
BTS v RI	Part 1	45.20	[11.89, 78.50]	2.67	.0079*
BTS+IP v RI	Part 1	11.96	[-14.22, 38.14]	0.90	.37
BTS v RI	Part 2	-7.46	[-37.56, 22.63]	-0.49	.63
BTS+IP v RI	Part 2	3.76	[-33.76, 41.29]	0.20	.84
BTS v RI	Total	36.79	[-15.05, 88.63]	1.39	.16
BTS+IP v RI	Total	15.07	[-37.31, 67.45]	0.57	.57

Note. * Significant results were determined using a Bonferroni-corrected threshold of α = .0083.

Visualisations

Figure B1

Proportional Distribution of Responses to All Survey Questions Across Conditions

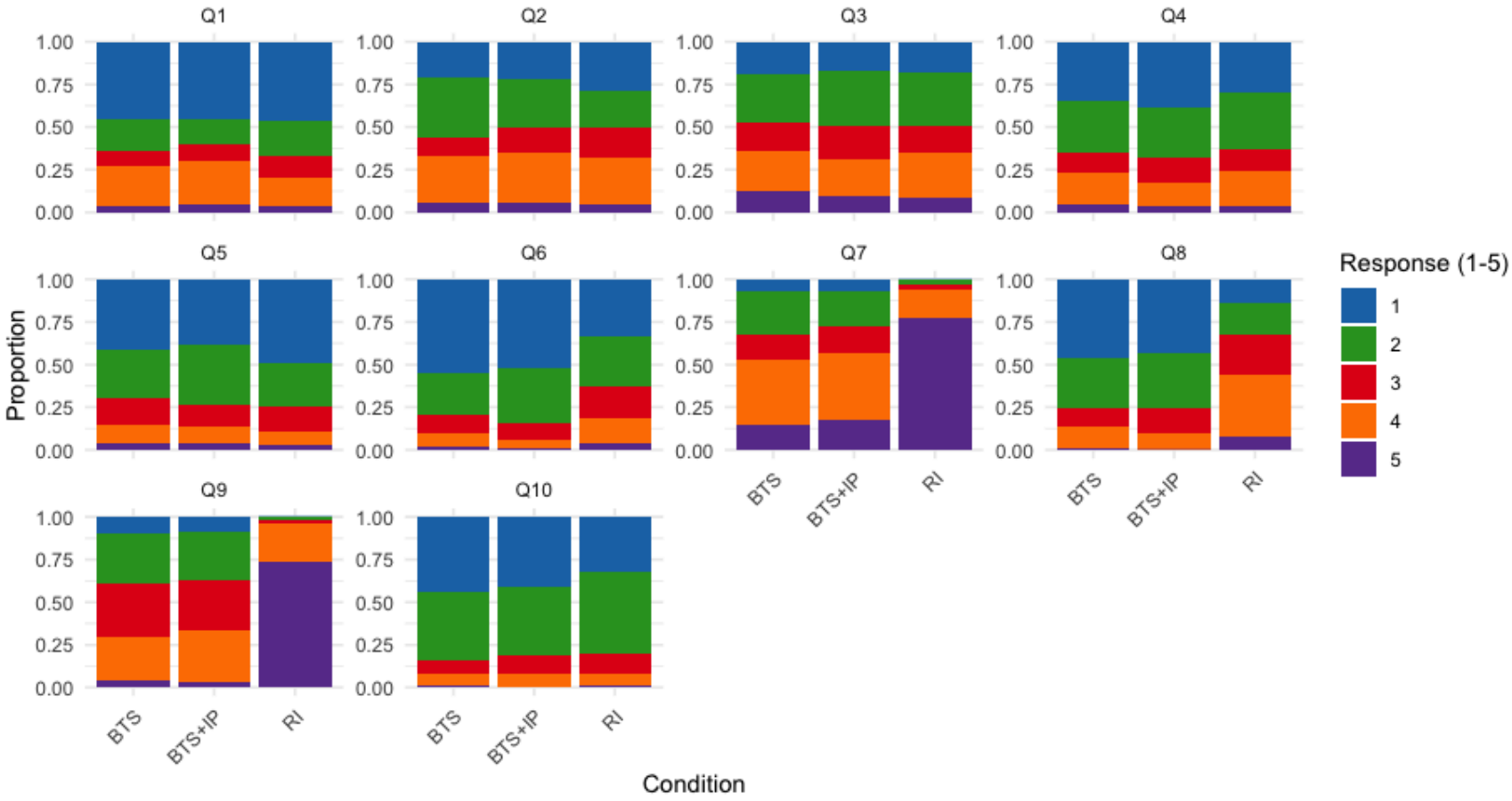


Figure B2

Differences in Median Responses For All Survey Questions Across Conditions