ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 365

The Social Desirability Atlas

Leonardo Bursztyn
Nicolas Röver
Ingar Haaland
Christopher Roth

May 2025

www.econtribute.de

UNIVERSITÄT BONN
UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

The Social Desirability Atlas*

Leonardo Bursztyn Ingar Haaland Nicolas Röver Christopher Roth

May 26, 2025

Abstract

Social desirability bias (SDB) is a pervasive threat to the validity of survey and experimental data. Respondents might often misreport sensitive attitudes and behaviors to appear more socially acceptable. We begin by synthesizing empirical evidence on the prevalence and magnitude of SDB across various domains, focusing on studies with individual-level benchmarks. We then critically assess commonly used strategies to mitigate SDB, highlighting how they can sometimes fail by creating confusion or inadvertently increasing perceived sensitivity. To help researchers navigate these challenges, we offer practical guidance on selecting the most suitable tools for different research contexts. Finally, we examine how SDB can distort treatment effects in experiments and discuss mitigation strategies.

Keywords: Social Desirability, Surveys, Experiments, Mitigation Strategies
JEL codes: B41, C83

*Leonardo Bursztyn, University of Chicago and NBER, email: bursztyn@uchicago.edu, Ingar Haaland, NHH Norwegian School of Economics, FAIR, CEPR, email: Ingar.Haaland@nhh.no; Nicolas Röver, University of Cologne, email: nicolas.roever@wiso.uni-koeln.de; Christopher Roth, University of Cologne and ECONtribute, NHH Norwegian School of Economics, Max Planck Institute for Research on Collective Goods, CEPR, email: roth@wiso.uni-koeln.de. We thank our discussant, Glenn Harrison, for extremely constructive comments that significantly improved the quality of the review. We also thank Luca Braghieri, Stefano Fiorin, Arkadev Ghosh, Luca Henkel, Lukas Hensel, Aaron Leonard, Matt Lowe, Michel Marechal, Anastasia Nebolsina, Ricardo Perez-Truglia, Krishna Srinivasan, and Andreas Stegmann for helpful feedback. Miguel Camacho Horvitz, Maximilian Fell, Milena Jessen, Luca Michels, and Max Müller provided excellent research assistance. Roth acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 and RCN through its CoE Scheme, FAIR project No 262675.

1 Introduction

Economists increasingly rely on self-reported data to measure outcomes such as income, financial decisions, welfare participation, and voting. A common concern with such data is that respondents might misreport their responses to appear more socially desirable. For instance, a welfare recipient might deny receiving government assistance, while a non-voter might still claim to have voted in the last election. This tendency, known as social desirability bias (SDB), arises when people align their answers with perceived social norms rather than reporting their true behaviors or views (Tourangeau and Yan, 2007). If left unaddressed, SDB can systematically distort survey estimates and treatment effects, undermining empirical conclusions. Quantifying the prevalence of SDB and identifying effective strategies to mitigate it are thus critical objectives for advancing the validity of survey and experimental research.

This review begins by defining SDB as a systematic bias in self-reports toward socially desirable answers. We argue that there are three distinct sources that can lead to SDB: (1) *material costs*, occurring when respondents fear material consequences from disclosing norm-violating behavior, such as legal repercussions; (2) *social-image concerns*, arising when individuals worry that their answers will lead to negative judgment by others; (3) *self-image concerns*, emerging when admitting certain attitudes or behaviors conflicts with a person's preferred self-image. These three forces may interact to determine whether and how SDB manifests in a given setting.

We then synthesize empirical evidence on its prevalence across a range of domains. Our analysis focuses on studies that combine self-reported data with individual-level ground-truth measures. This criterion serves to restrict the sample to high-quality studies and allows us to test whether we can observe over- and underreporting patterns suggested by SDB on an individual level. Our data are consistent with SDB—and not random measurement error—severely distorting self-reported responses.

To identify where SDB is most prevalent, we compare direct question estimates and validation data from secondary sources across six domains: criminal behavior, economic outcomes, educational outcomes, health behaviors, moral behavior and voting. We document that the influence of SDB is heterogeneous, but can be substantial. In the context of business tax evasion, social security fraud and cheating, self-reports are more than 50% lower than the actual number retrieved from secondary sources (Höglinger and Jann, 2018; Markhof et al., 2025; van der Heijden et al., 2000).

Having established that SDB is a severe concern across many domains, we then critically assess commonly used strategies to mitigate SDB. These include list experiments (LEs), randomized response techniques (RRTs), anonymity guarantees, social circle questions, forgiving outcome framing, vignette experiments, and the use of incentives. We highlight how some strategies can backfire by creating confusion or inadvertently increasing perceived sensitivity of the item at hand. In particular, the evidence using individual-level benchmarks reveals that commonly used mitigation strategies—such as LEs and RRTs—sometimes fail to improve survey data accuracy. Instead, the complexity of these indirect-question formats can confuse respondents or unintentionally increase the perceived sensitivity of the topic, undermining their intended benefits (John et al., 2018; Loewenstein, 1999; Markhof et al., 2025).[1]

To help researchers navigate these challenges, we offer practical guidance on selecting the most suitable tools for different research contexts. The effectiveness of any strategy to mitigate SDB depends critically on understanding the underlying motivations behind respondents' misreporting: material costs, social-image concerns, and self-image concerns. When respondents fear material costs—such as punishment or institutional backlash in sensitive contexts—researchers should use those strategies that offer the strongest anonymity protections. Second, when social-image concerns predominate, privacy-enhancing methods like LEs are particularly effective by obscuring individual identities. Finally, self-image concerns can potentially be addressed with techniques such as third—person framing or forgiving outcome framing. Consequently, effectively addressing SDB requires first identifying the primary motives for misreporting and then selecting a mitigation strategy accordingly.

While our discussion so far has focused on how SDB distorts estimates of the prevalence of behaviors such as voting and welfare claims, an important concern for economists is also whether—and under what circumstances—SDB distorts treatment effects in experiments. Building on de Quidt et al. (2018), we clarify that this issue arises in experimental settings where treatments alter the salience and perceptions of what constitutes a socially desirable answer. We propose a set of best practices designed to mitigate differential SDB while preserving statistical power.

[1]The finding that the complexity of these methods may substantially increase confusion is in line with a growing literature on responses to complexity (Enke and Graeber, 2023).

Related Literature. This review connects with a vast literature on SDB in psychology, sociology, and political science. Early work by Edwards (1957) and Crowne and Marlowe (1960) viewed social desirability as a *stable individual trait*, which can be captured by psychometric scales—such as the Marlowe-Crowne Social Desirability Scale. Seminal texts on survey methodology by Sudman and Bradburn (1974) and Bradburn (1978) subsequently provided evidence that situational factors, such as question wording, interview mode, and interviewer–respondent dynamics, systematically shape answers to sensitive questions. The importance of situational factors has been further explored in more recent work (Blair et al., 2020; Tourangeau and Yan, 2007; Yan, 2021). Building on this previous work, our review focuses on applications and methods prominent in economics and is, to our knowledge, the first to provide an overview of misreporting using studies that rely on individual-level ground-truth measures.

We also relate to reviews on the design of surveys and experiments in economics (e.g., Fuster and Zafar, 2023; Haaland et al., 2023, 2025; Harrison and Swarthout, 2025; Harrison and List, 2004; Stantcheva, 2023) and articles discussing the validity of experimental and survey data (e.g., Bertrand and Mullainathan, 2001; Levitt and List, 2007). Compared to these reviews, we provide a more in-depth treatment of SDB and an empirical assessment of the prevalence of SDB across domains.[2]

Finally, our work builds on a literature studying gaps in private and public behavior to identify social image concerns (Bursztyn and Jensen, 2017). This literature typically relies on experimental manipulations of visibility. Empirical evidence confirms large private-public gaps across a variety of domains (Ajzenman et al., 2024; Braghieri, 2024; Bursztyn et al., 2020b, 2023b; DellaVigna et al., 2017).[3] Our review focuses on distortions from SDB arising even in settings that are plausibly anonymous, such as online surveys.

This review proceeds as follows. Section 2 provides a simple definition of SDB and discusses different motives underlying misreporting. Section 3 synthesizes empirical evidence on the prevalence and magnitude of SDB using ground-truth benchmarks. Section 4 assesses mitigation strategies and discusses evidence on their effectiveness. Section 5 makes recommendations for matching different mitigation strategies to the research context. Section 6 explains how SDB can

[2]Our review also relates to work studying experimenter demand effects, which occur when respondents infer the researcher's expectations and adjust their answers accordingly (e.g., de Quidt et al., 2018, 2025). Perceived experimenter expectations may diverge from what is considered the socially desirable response.

[3]Responses given in public conditions can be interpreted as an upper bound for socially desirable responding. Private responses, however, may still lie well above a lower bound—particularly when self-image concerns are at play.

bias treatment effects and reviews strategies for mitigation. Section 7 concludes.

2 Conceptualizing Social Desirability Bias

2.1 Definition

SDB is the systematic gap between a respondent i's latent or "true" attitude y_i^* and the answer y_i they give when they expect that answer to be evaluated against a perceived *social norm* N_i (Paulhus, 1984; Tourangeau and Yan, 2007). Here, we define N_i as a first–order normative expectation: the response that the respondent believes a typical or relevant member of her reference group *thinks* one *should* give in the same situation (cf. Bicchieri, 2006). Formally, N_i is the respondent's subjective estimate of the modal "ought" response. Respondents trade off the cost of violating N_i against the disutility of misreporting. When $N_i \neq y_i^*$, this trade-off may produce a directional bias.

2.2 Why Do People Respond in Socially Desirable Ways?

We distinguish three distinct sources of SDB which we discuss in detail below.

Material Costs. One source of SDB arises when respondents perceive tangible costs to disclosing norm-deviating attitudes or behaviors (Blair et al., 2020). When revealing a behavior could expose a respondent to legal, financial, or other tangible consequences, they have a strong incentive to misreport or conceal the truth. For instance, admitting to drug use or benefit fraud might carry legal repercussions or lead to a loss of welfare eligibility. Similarly, disclosing past criminal activity, workplace misconduct, or politically sensitive views could jeopardize employment prospects. In some cases, respondents may fear that revealing unauthorized work or visa violations could endanger their immigration status. Respondents may also tailor their responses to increase their chances of receiving future benefits or maintaining relationships with the researchers.

Material costs become more relevant the higher the chances that survey responses are not fully anonymous. While surveys are increasingly conducted using online labor markets, even survey takers who only provide their researchers with a seemingly anonymous ID might still not trust that their anonymity is fully protected. For instance, Lease et al. (2013) show that the seemingly

anonymous worker ID on the platform Amazon Mechanical Turk was, in fact, not anonymous and could reveal personally identifying information. While platforms such as Prolific claim that the IDs that uniquely identify workers are truly anonymous, workers on the platform might still worry that a security breach could expose their identities. Furthermore, by providing researchers with rich demographic data, they could potentially identify themselves. Sweeney (2000) shows that only three demographic variables (5-digit ZIP, gender, date of birth) can uniquely identify 87% of the US population, which could render this a rational fear, especially in light of the widespread sharing of research data in public research repositories (which typically include the demographic data). Thus, it might seem prudent for respondents not to reveal sensitive or illegal behaviors in surveys.

Social Image Concerns. Respondents may distort survey answers due to social image concerns when anonymity is uncertain, as perceived reputational risks incentivize them to align responses with social norms (e.g., Bursztyn and Jensen, 2017; Cortés et al., 2024; Kuran, 1998). Concerns about social stigma or judgment may motivate respondents to underreport potentially stigmatized behaviors (Bursztyn et al., 2023b) and exaggerate virtuous traits (Ewers and Zimmermann, 2015).

Self-Image Concerns. Even if respondents are certain that their responses are anonymous, self-image concerns may significantly drive socially desirable responses (e.g., Bénabou and Tirole, 2002, 2016; Henkel et al., 2024). For example, individuals may experience discomfort admitting outcomes, behaviors, or attitudes that conflict with their preferred self-image (e.g., Bénabou and Tirole, 2002). To maintain a positive self-image, respondents may engage in various forms of motivated reasoning and cognition, such as holding self-serving beliefs about others (e.g., Di Tella et al., 2015) and selectively attending to information during belief updating (e.g., Jiao, 2020) or retrieval (e.g., Zimmermann, 2020).

3 Severity of SDB Across Domains

In this section, we examine misreporting patterns to determine whether and to what extent SDB distorts estimates based on self-reports. We first show strong evidence that SDB is an important source of errors in self-reports (Subsection 3.1). Then, we present data on SDB's influence

across a wide variety of domains (Subsection 3.2).

Our analyses draw on studies that compare multiple elicitation techniques, including the direct question format, and verify results using *individual-level* ground-truth measures. We make individual-level validation a core requirement for two reasons. First, studies that compare survey estimates only to aggregate benchmarks can be misleading, especially if the survey population systematically deviates from the population of interest. Second, without individual-level benchmarks there is an identification problem: in that case the researcher cannot disentangle false positives from true positives and false negatives from true negatives. Only individual-level validation data reveals the fraction of respondents that actually engaged in socially undesirable behaviors that misreported it. We summarize all studies that meet our inclusion criteria in Appendix Table A1.[4]

3.1 Does SDB Decrease Accuracy of Self-Reports?

To present evidence that SDB systematically distorts survey responses, we now compare false positive rates (FPRs) and false negative rates (FNRs) across a subset of studies which report these statistics.

The FPR is calculated over the subset of respondents who have truly *not* performed the behavior of interest; it is defined as the share of these individuals who self-report having performed the behavior. For instance if the behavior is tax evasion, the FPR is the share of respondents who report tax evasion among all honest taxpayers. In turn, the FNR is calculated over the subset of respondents where external records show they performed a behavior; it equals the share of these respondents who deny the behavior. In the tax-evasion example, the FNR is the share of people who report being a compliant tax payer among all true tax evaders.[5]

The presence of SDB implies systematic differences between these two rates: If a behavior is socially undesirable, people who have performed it are prone to deny (leading to a high FNR)—many tax-evaders do not report evading taxes. Conversely, non-performers seldom misreport (leading to a low FPR)—few compliant taxpayers report having evaded taxes.

Figure 1 presents our data, which is very consistent with SDB causing a majority of response

[4]To increase coverage of topics for this section, we compile a list of additional high-quality studies comparing direct question estimates and individual-level ground-truth data from secondary sources. These studies are listed in Appendix Table A2.

[5]See Appendix Section D for more detailed explanation of FPRs and FNRs.

errors.[6]

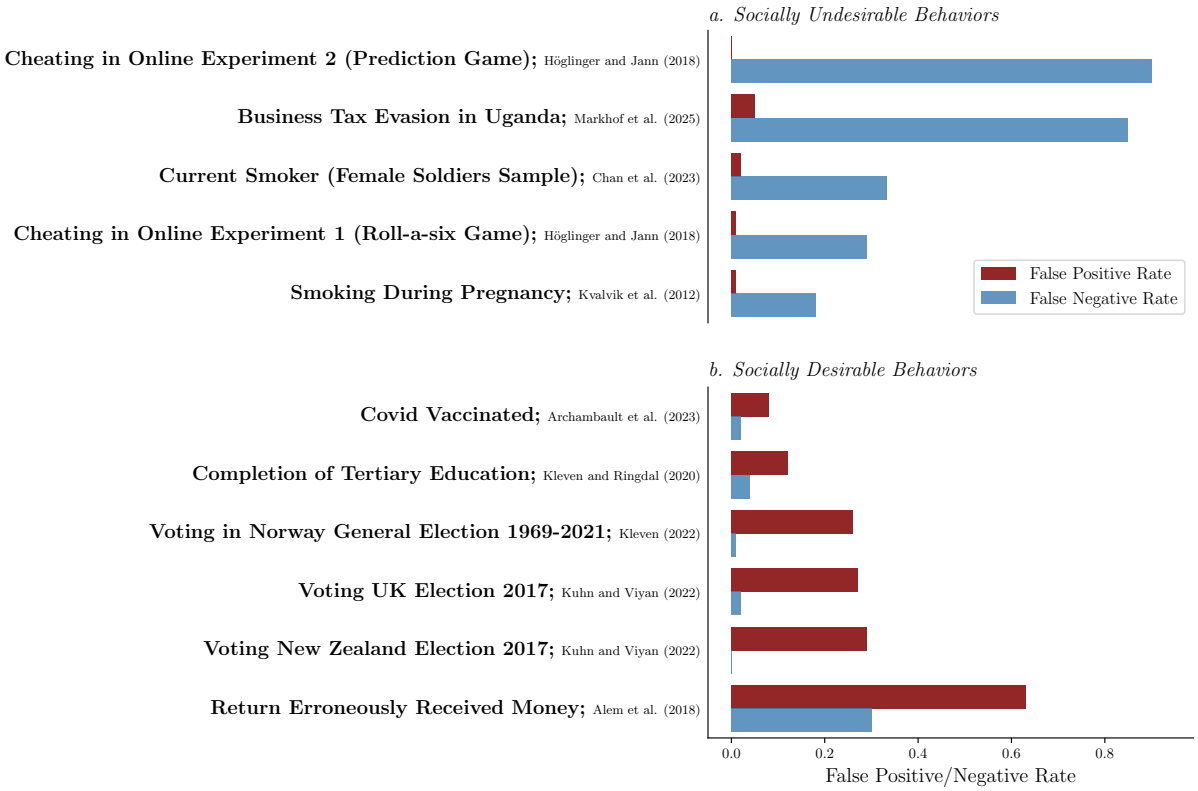


Figure 1: False-negative and false-positive rates by desirability of the behavior from direct-question self-reports. For the underlying data, see Appendix Table A3.

For socially undesirable behaviors, FNRs are several orders of magnitude larger than FPRs; this dynamic reverses for socially desirable behaviors. And in certain domains, error magnitudes are so large that estimates become completely unreliable. For tax evasion, the false negative rate is 85%, meaning that on average 85 out of 100 tax-evaders remain undetected from self-reports (Markhof et al., 2025).

3.2 Which Topics are Prone to SDB?

Having established that SDB can be a source of severe bias, we now turn to creating an atlas that maps its influence across diverse behavioral domains. We display the prevalence of a given

[6] We choose to use FPRs and FNRs for this exercise as this is the most consistent way to compare our studies. However, the FPR *and* and FNR are not available for many studies for several reasons. First, some studies focus on a sample where all subjects have engaged in the behavior, e.g. all subjects are convicts, meaning by definition there cannot be false positives. Second, FPRs and FNRs are only defined for binary outcomes. Third, some studies have a different focus, and simply do not report both measures.

behavior estimated using direct survey questions divided by the ground-truth prevalence in Figure 2. The figure reports the ratio across six domains: criminal behavior, economic outcomes, educational outcomes, health behaviors, moral behavior, and voting. Values lower than one are indicative of underreporting, while values higher than one indicate overreporting.

Taken together, this way of presenting the data strongly suggests an influence of SDB as well: Respondents systematically underreport behaviors viewed as undesirable and overreport those viewed as desirable. We now discuss every domain in detail.

Criminal Behavior In this domain, underreporting is substantial and the accuracy of direct question estimates is low. This may reflect that these topics likely activate all three channels motivating strategic misreporting as discussed in Section 2. Respondents face not only potential material consequences (e.g., legal or financial penalties if tax evasion becomes public), but also heightened self-image and social-image concerns, given the strong stigma associated with criminal activity.

Economic Outcomes Outcomes in the economics domain are subject to systematic misreporting to varying degrees. On average, SDB is comparable in severity to other domains, such as moral behavior and voting. Substantial underreporting is documented for bankruptcy (Locander et al., 1976), whereas underreporting related to unemployment insurance applications appears more moderate (Dutz et al., 2021; Kirchner, 2015).

Education In the domain of education, respondents underreport having ever failed a class but overreport having completed some form of tertiary education (Kleven, 2022; Lamb and Stem, 1978). However, error magnitudes for these outcomes are quite small and accuracy of the direct question estimates is comparatively high.

Moral Behavior Survey items touching on honesty of respondents are very prone to SDB as honesty is an important virtue in almost all societies (Cohn et al., 2019). Both John et al. (2018) and Höglinger and Jann (2018) conduct experiments where subjects can cheat and then ask them to self-report whether they actually cheated. Both document large discrepancies between reported and actual cheating in some of their settings.

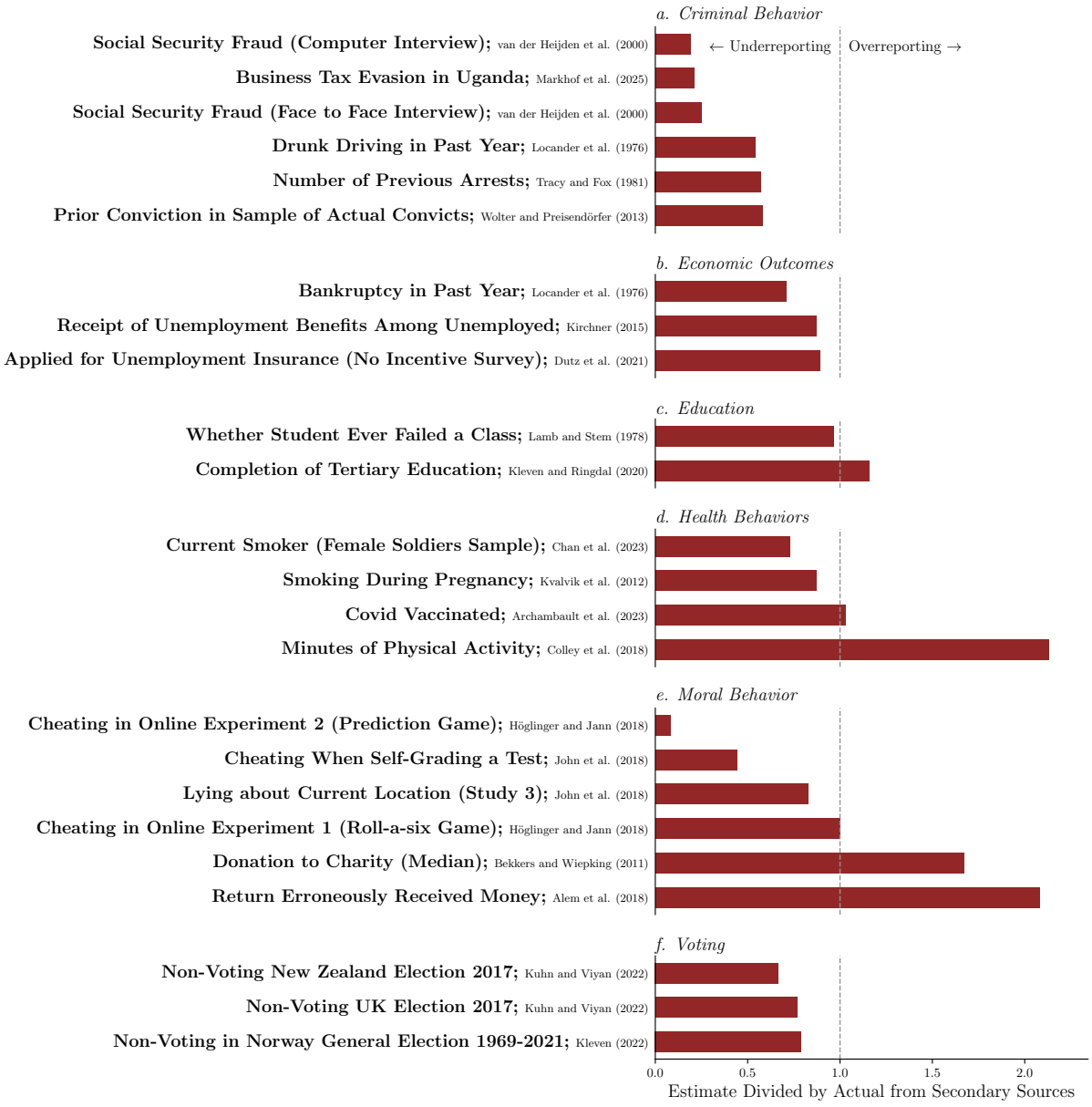


Figure 2: Estimates of the prevalence derived from the direct question self-reports divided by actual prevalence obtained from secondary sources.

Health Behaviors Substantial biases emerge prominently in health-related behaviors, including pronounced underreporting of physical inactivity (Colley et al., 2018), smoking during pregnancy (Kvalvik et al., 2012), and current smoking status (Chan et al., 2023). These patterns likely reflect intense social pressures around responsible health behaviors and personal accountability.

Voting Finally, several studies show that SDB distorts self-reported voting estimates. In democratic societies, voting is widely regarded as a civic duty (DellaVigna et al., 2017), which means both self-image and social-image concerns are present when individuals respond to questions about their voting behavior. Supporting this interpretation, Karp and Brockington (2005) show that misreporting is systematically related to the salience of the social norm to vote. Kleven (2022) exploits five decades of Norwegian election surveys individually linked to administrative turnout records and shows a 96% agreement between survey answers and register data—higher than in most comparable studies. Consistent with SDB, he finds that overreporting is far more common than underreporting. Furthermore, while the survey-registry agreement is relatively high overall, it is considerably lower for certain subgroups, such as younger respondents with low education. In an experiment during the 2019 local-election, Kleven and Bergseteren (2023) randomly assigned respondents to phone or web interviews. Consistent with heightened social pressure in more personal interactions, they observe significantly higher overreporting in the phone condition (4.2%) than in the web condition (2.8%), while underreporting stayed at 0.3% in both modes.

To conclude this section, there is evidence for SDB in a series of other domains for which no individual-level benchmarks are available, including sexual behavior (Björkman Nyqvist et al., 2018), racial or xenophobic attitudes (Hainmueller and Hangartner, 2013; Kuklinski et al., 1997), and corruption (Kraay and Murrell, 2016).

4 Mitigating Social Desirability Bias

While the previous section measures the extent of SDB across domains, this section turns to techniques explicitly designed to mitigate SDB. The approaches covered include list experiments, randomized response techniques, explicit anonymity guarantees, social circle questions, forgiving outcome framing, vignette experiments, and monetary incentives.

4.1 List Experiments

The list experiment (LE), also known as the item count technique, estimates the prevalence of a sensitive outcome by indirectly inferring its presence. The LE is widely used in economics research to measure preferences and behaviors that are subject to SDB (e.g., Bursztyn et al., 2020a; Chen and Yang, 2019; List, 2025).[7]

Procedures Respondents are randomly assigned to a control group, which receives a list of non-sensitive items, or a treatment group, which sees the same list with an additional sensitive item (Miller, 1984).

The intuition becomes clear in an example. Suppose the control group is given a list of three items—liking coffee, owning a pet, and having traveled abroad—while the treatment group sees the same three items plus a fourth: having used illicit drugs. If the average number of items endorsed in the control group is 1.8 and in the treatment group is 2.3, then the difference, 0.5, estimates the share of respondents who have used illicit drugs.

The logic behind this method is that it removes the direct link between an individual's response and the sensitive item. If someone reports endorsing three items, there is no way to know whether that includes the sensitive item or not for that specific individual. In other words, individuals have plausible deniability, which should reduce material concerns and social-image concerns.

Double LEs extend traditional LEs by giving each respondent two separate lists, each containing several non-sensitive items, with the sensitive item randomly placed on one of the lists (Glynn, 2013). This design exploits both within-subject and between-subject variation to enhance statistical efficiency and reduce standard errors compared to standard single LEs.

Assumptions Two assumptions must hold for the LEs to yield unbiased estimates. First, adding the sensitive item must not change how respondents answer the non-sensitive items (Coffman et al., 2017). Second, respondents must report the total count truthfully (including the sensitive item only if it applies). If respondents withhold or mistakenly add the sensitive item (e.g., because of the increased complexity of the elicitation), the method will fail to uncover the true prevalence of the sensitive item.

[7] For a primer on the statistical analysis of LEs, see Blair and Imai (2012)

Best Practices The design of LEs should try to minimize participant confusion to ensure reliable responses. Keeping lists short (3–5 items) reduces cognitive burden and miscounting errors, while control items should be neutral and contextually similar to the sensitive item to avoid distortion. Clear, simple instructions with a practice example improve comprehension. Pilot testing helps refine wording and detect misunderstandings. Using negatively correlated statements reduces full-agreement patterns that compromise anonymity (Coffman et al., 2017) and item order should be randomized to prevent biases.

To enhance accuracy, non-sensitive items should have low response variance (Coffman et al., 2017). When comprehension is a concern, consistency tests—asking respondents two lists with the same sensitive item but different controls—help assess compliance (Chuang et al., 2021). However, this approach comes at the potential cost of increasing the salience of the sensitive item.

Potential Pitfalls and Limitations There are four potentially important limitations with this method. First, because the LE does not hide answers from the respondents themselves, the method will not reduce sensitivity biases arising from self-image concerns. Second, a potential limitation with interventions that saliently provide anonymity, like the list method, is that it might actually increase the salience of the sensitivity of the question (Loewenstein, 1999). This could result in underreporting of the sensitive item even within the list format. Third, participants may also struggle with the format of the list method, miscounting items or failing to understand the task, which introduces noise. Fourth, LE estimators have a larger variance than direct questions, requiring large sample sizes.

Evidence on Effectiveness Several studies in the social sciences have compared LEs to direct questioning and find that LEs tend to produce higher estimates of sensitive behaviors compared to direct questioning on average (Blair et al., 2020; Ehler et al., 2021; Li and Van den Noortgate, 2022). However, these studies do not rely on individual level ground-truth benchmarks and hence cannot provide insights whether the increases in prevalence are driven by increases in true positives (i.e. individuals engaging in socially undesirable behaviors admitting to it) or false positives (i.e. individuals not engaging in socially undesirable behaviors wrongly admitting to it).

To our knowledge, only two studies compare LE estimates to individual-level ground-truth

benchmarks, and both highlight potential limitations of the method. Markhof et al. (2025) examine tax evasion among Ugandan firms and find that the list method significantly reduces estimation accuracy. Specifically, they report a false positive rate of approximately 18%—compared to just 5% when using direct questions.[8] The authors attribute this discrepancy to the complexity of the list task and respondent misunderstanding. Similarly, in the context of voter turnout, Kuhn and Vivyan (2022) find that the list method produces significantly less accurate estimates than direct questioning. They provide evidence that this is driven by respondents providing a reasonable number without fully engaging with the task.

4.2 Randomized Response Technique (RRT)

Unlike the LE, which masks individual responses through aggregation, the RRT offers more privacy through randomization.

Procedure Pioneered by Warner (1965), the RRT introduces a probabilistic element to responses. Instead of directly answering a sensitive question, respondents follow a randomization procedure—such as flipping a coin in private—to determine whether they answer truthfully or provide a predetermined response. Since the researcher does not know which question is answered, plausible deniability is preserved, reducing material and social-image concerns.

For example, to measure illicit drug use, respondents are instructed to flip a coin: if heads, they answer "Have you ever used illegal drugs?"; if tails, respondents are instructed to simply say "yes". If 20 out of 100 respondents answer "no", then the estimate for illicit drug use is $\frac{30}{50} = 60\%$, because half of the sample were forced to say "yes" from the tails outcome.

A different version of the RRT is the crosswise method (Yu et al., 2008), which increases statistical power compared to the standard RRT. Here, respondents never directly answer the sensitive question; instead, they indicate whether their answers to two separate yes/no questions (one sensitive, one non-sensitive) are identical or different. However, compared to the standard RRT, the crosswise technique requires respondents to understand slightly more complex instructions, which could pose comprehension challenges.

[8]The false positive rate is defined here as the number of firms falsely reporting tax evasion divided by all firms reporting tax evasion. See Appendix Table A3 and Appendix Section D.

Assumptions The RRT assumes that respondents faithfully follow the randomization protocol, and that respondents comply to answering truthfully. Also, the randomization outcome must be independent of respondents' true responses. If participants can anticipate or influence the outcome, privacy is compromised.

Potential Pitfalls and Limitations The RRT is subject to similar pitfalls and limitations as the LE. It cannot resolve biases arising from self-image concerns, involves a rather complex elicitation procedure and very saliently provides anonymity—which can ring alarm bells in subjects (John et al., 2018). Coutts and Jann (2011) suggests that respondent trust issues are even more pronounced for the RRT than for LEs: In their experiments, subjects report low levels of trust that the RRT provides complete anonymity (below 23% in all RRT conditions), while at the same time reporting they had completely understood the instructions (above 79% in all RRT conditions). As with the LE, an implementation of the RRT requires relatively large sample sizes.

Best Practices Practice rounds help minimize confusion, allowing respondents to get familiar with the randomization before answering the items of interest (Rosenfeld et al., 2016). Without this, misinterpretations can introduce systematic bias. Trust is just as important—if respondents suspect their true answer might still be inferred, they may default to socially desirable responses, undermining the method's purpose. Reinforcing anonymity is essential to securing compliance—especially in light of the results from Coutts and Jann (2011). For a more in-depth coverage of best practices in implementation, we refer to Blair et al. (2015).

Evidence on Effectiveness A meta-analysis by Lensvelt-Mulders et al. (2005) find that RRT on average increases the reported prevalence of stigmatized behaviors by 11 percentage points. Rosenfeld et al. (2016) shows that the RRT increases the estimated prevalence of sensitive behaviors compared to both direct questioning and the list method. Boudreau et al. (2023) document higher reporting of harassment using the RRT compared to a direct question. Yet, these studies do not rely on individual-level ground-truth measures. As for the LE, this means that increases in estimates of prevalence could be driven by false positives, e.g. from non-strategic reporting due to confusion.

To provide more conclusive evidence on the effectiveness of the RRT, we focus on studies with

individual-level benchmarks. Figure 3 plots the ratio of RRT-based estimate of the prevalence of a behavior and the actual prevalence. The figure highlights substantial variability in the ratio of the RRT-based estimate and the actual prevalence. This evidence suggests that the RRT performs poorly in many settings.

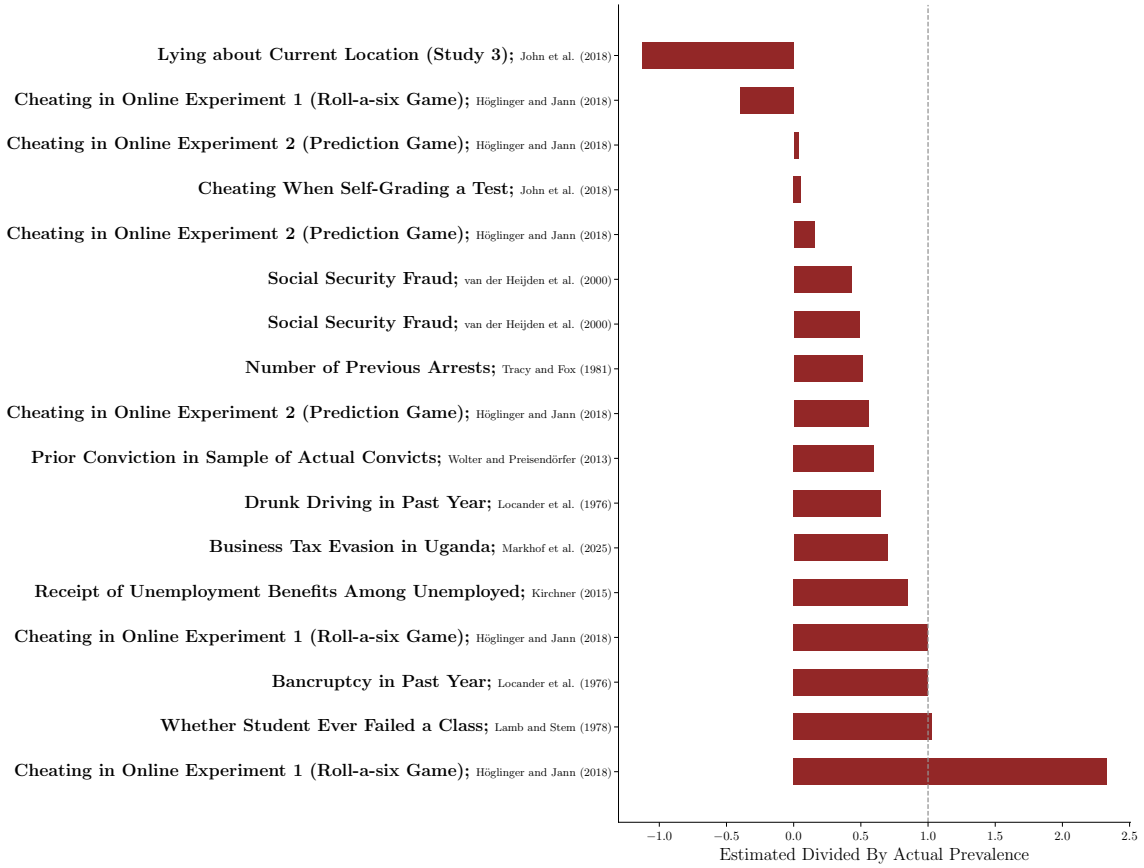


Figure 3: Ratio of RRT-estimates and actual prevalence based on studies with individual-level benchmarks. In case the same reference appears multiple times, the paper reports results from various variations of the RRT. The RRT can provide non-sensical, negative estimates if people who have to answer the non-sensitive question give the wrong answer (e.g. if the RRT instructs the "control group" to answer "yes" if they are born in January/February and people who are born in these months answer "no").

Two main explanations have been proposed for this limited performance—both of which may also apply to other indirect questioning methods, such as LEs. First, John et al. (2018) provide compelling evidence that the RRT increases the perceived sensitivity of the question. This heightened salience may prompt respondents to become more protective of their information, resulting in systematic underreporting. Second, another series of papers highlights that the complexity of the RRT introduces non-classical measurement error, which could bias the estimate in either way (e.g. Höglinger and Diekmann 2017; Markhof et al. 2025).[9]

[9]Non-compliance with instructions among surveyed individuals appears prevalent and not easy to characterize

4.3 Anonymity Guarantees

Anonymity guarantees are a common method to reduce SDB by reassuring respondents that their answers are not traceable, aiming to reduce material and social-image concerns.

Procedure Common practices include administering surveys in self-paced, online formats that do not collect identifying information, and clearly communicating that responses will remain confidential.

Assumptions and Potential Pitfalls For anonymity guarantees to reduce SDB respondents must trust these guarantess. However, using such guarantees has potential pitfalls that may attenuate their intended effects. In lab and field experiments, ensuring full anonymity may reduce engagement or alter behavior in unintended ways. For example, in dictator games, removing social pressure may not only reduce giving but also encourage random or unusual decision-making (Hoffman et al., 1996). Excessive emphasis on anonymity might signal that a question is particularly sensitive, potentially increasing item non-response or generating defensive answers (Loewenstein, 1999). It might also lead to less truthful reporting by priming respondents to worry that their answers are not truly anonymous. Indeed, one way to think of anonymity guarantees is to view them as "cheap talk" (Farrell and Rabin, 1996).

Best Practices Effectively reducing SDB through anonymity guarantees requires careful implementation. Avoiding emphasis on anonymity that respondents might interpret as signaling question sensitivity can further mitigate defensive answering. Attention checks and pre-tests can assess whether anonymity mechanisms function as intended. Additionally, online surveys appear less prone to SDB than in-person or telephone-administered surveys (Holbrook and Krosnick, 2010; Kleven and Bergseteren, 2023; Reisinger, 2022).

Evidence on Effectiveness Evidence on how privacy affects survey outcomes remains limited. Respondents appear more willing to admit to behaviors such as drug use or tax evasion when assured anonymity (Tourangeau and Yan, 2007). Barmettler et al. (2012) investigate the effect of experimenter-subject anonymity on decisions in dictator, ultimatum and trust games. They find no significant effect of their anonymity manipulations.

(Chuang et al., 2021). See also Appendix Table A3 for evidence on false positives and negatives.

4.4 Social-Circle and Third-Person Questions

Social-circle and third-person questions aim to reduce SDB by changing the target of evaluation, allowing researchers to infer an individual's attitude or behavior from that individual's perceptions about others (Haire, 1950; Ling and Imas, 2025).

Procedure. These question formats have been applied in different ways. For example, Galesic et al. (2018) ask respondents to report on the voting intentions of their social circle rather than their own, and find that such aggregate perceptions yield more accurate predictions of election outcomes. Another approach involves systematically varying the reference group—for instance, asking about the prevalence of a sensitive behavior among in-group versus out-group members. This variation helps identify whether SDB is present, as respondents may be more susceptible to self-image and social-image concerns when considering close social circles (Chakravarty et al., 2022). Bursztyn et al. (2020a) and Bursztyn et al. (2023a) employ a related method. Some participants are asked about the extent to which they believe that others "would say that they agree" with a policy, while other participants are asked about the extent to which they believed that others "would truly agree" with the given policy. They document muted differences in the answer distributions, from which they infer a limited importance of SDB in their settings.

Assumptions and Potential Pitfalls Because these approaches rely on beliefs about others—i.e., higher-order beliefs—they place greater cognitive demands on respondents than direct self-reports. Social circle questions work well when social networks are relatively homogeneous and the social circle is well defined and familiar to the survey respondent. When respondents draw from heterogeneous or poorly defined social circles, the resulting data can suffer from substantial noise and measurement error. Additionally, empirical research documents pluralistic ignorance in some contexts—systematic misperceptions about the behaviors or attitudes of peers, which further limits the reliability of social circle questions in these settings (e.g., Bursztyn et al., 2020a).

Best Practices When implementing social-circle questions, researchers should carefully define the reference group to ensure clarity and consistency in interpretation. Ambiguity about whether "peers" refers to friends, coworkers, neighbors, or a broader social category can introduce noise and reduce the validity of the measure. To improve comparability, it is useful to explicitly specify

the group (e.g., "your close friends" or "people your age in your community") and keep this consistent across treatments. The framing of the question should also be neutral and avoid priming respondents toward particular norms. Finally, assessing the plausibility of projection in the specific context—e.g., by measuring perceived similarity with peers—can help researchers gauge the reliability of inferences drawn from these responses.

Evidence on Effectiveness Evidence on effectiveness is rather scarce and does not allow for a clear evaluation of the method's effectiveness. (Fisher, 1993) show that the method often reveals higher prevalence of sensitive. Galesic et al. (2018) provide encouraging evidence in the context of estimating election results, and Bursztyn et al. (2020a) and Bursztyn et al. (2023a) use their method to argue that SDB is not present in their settings.

4.5 Forgiving Outcome Framing

Forgiving outcome framing offers a pragmatic, low-cost way to curb SDB when self-reports are unavoidable. The core idea is to embed questions in a brief normalizing narrative which signals that a broad range of behaviors is common and acceptable. For example, instead of asking respondents whether they *always* comply with public-health guidelines, researchers might preface the question with: "People vary in how closely they follow public-health recommendations," before requesting a concrete frequency report for the past week. By explicitly acknowledging behavioral heterogeneity up front, the wording reduces normative pressure and makes deviations from the ideal less stigmatizing. Alternatively, socially undesirable behavior may be justified, for instance by attributing it to external factors.

Procedure Several strategies can be employed to make language more forgiving. First, researchers can introduce a brief normalizing statement acknowledging natural variation in the target behavior. Second, they can attribute socially undesirable responses to neutral external factors—such as workplace culture or limited time—to mitigate moral judgment. Finally, presenting a balanced response scale (e.g., "never," "rarely," "sometimes," "often,", "always") ensures that less desirable answers do not stand out as exceptional.

Assumptions and Pitfalls The approach rests on the assumption that respondents are more candid when a question protects their self-esteem by easing concerns about external judgment. Yet, forgiving wording can backfire. A preamble that is too explicit or morally charged may increase respondents' awareness of the sensitive nature of the item and thus emphasize rather than diminish bias (de Quidt et al., 2018). Finally, by signaling that shortcomings are widespread, forgiving outcome framing in itself could induce a demand effect (de Quidt et al., 2018).

Best Practices Several design principles can mitigate these risks. The preamble should be brief—preferably fewer than twenty-five words—to avoid spotlighting the sensitive issue. Moralistic adjectives such as "responsible" or "good" are best omitted in the question so that the SDB-reducing effect of the introduction is not offset by value-laden language.

Evidence on Effectiveness Empirical evidence on the efficacy of this technique is mixed. Belli et al. (2006) finds modest improvements in the accuracy of self-reported turnout when turnout questions have face-saving response options. Examining different question wordings examining sexual behavior, Catania et al. (1996) presents evidence that "supportive" wording results in increased reporting of sexual behaviors typically considered socially sensitive or taboo. Peter and Valkenburg (2011) study a similar context, the use of sexual media content. While they find no overall effect of a forgiving introduction, they document heterogeneity with regards to how respondents score on a social desirability scale. Among respondents with high scores on the scale, a forgiving introduction increases reports of consuming sexual media content. Studying four sensitive domains in an online survey, Näher and Krumpal (2012) find no consistent differences between a condition where respondents answer a simple direct question, and three other conditions where the questions are framed in a forgiving language or embedded in a more permissive/restrictive context.[10] The risks of employing forgiving framing are illustrated by Kaplan and Yu (2015). They uncover an increase in perceived sensitivity—especially among interviewers—when vignettes are framed in a forgiving way.

[10]Näher and Krumpal (2012) consider the domains election participation, cheating on a partner, driving under influence and use of antidepressants.

4.6 Vignette Experiments

Experiments with hypothetical vignettes can be a powerful way to mitigate SDB. Because the object of interest is embedded within a hypothetical scenario, where potentially many variables are simultaneously randomized, respondents may pay less explicit attention to the sensitive issue, thereby reducing social desirability concerns compared to direct questioning (Alexander and Becker, 1978). To illustrate this point, imagine investigating whether there is a "mother penalty" for prioritizing career over family. You could formulate the following hypothetical vignette:

> Marie works as an equity analyst at a large investment bank in Germany. She has two small children, aged 2 and 4. When her employer asks if she would like to spend 3 months in Singapore to work on an exciting project that could significantly increase her chances of promotion, she decides to accept the offer, leaving her children in Germany with their father. How acceptable do you find her decision to leave her children for 3 months?

By randomizing whether respondents see a mother or a father making this career decision, researchers can assess whether there is a "mother penalty" without explicitly prompting respondents to compare genders. In contrast, direct questioning might explicitly ask: "Would you disapprove more if a mother prioritized her career over her small children compared to if a father did the same?" Here, a combination of self-image and social-image concerns might make respondents much more likely to claim that they would not disapprove more of mothers prioritizing their careers.

Chopra et al. (2024) provide a recent example in economics on how vignette experiments can be a powerful tool to address social desirability concerns. They give rich descriptions of hypothetical academic studies, including a summary of the study design and the main results, and examine whether there is a "null result penalty" by varying whether the main result is statistically significant or not, keeping statistical precision identical across vignettes. They find a significant null result penalty even among academic editors who might have preferred reporting not to discriminate against null results if asked directly.

In vignette experiments, it is good practice to randomize multiple attributes of the vignettes. For instance, in addition to randomizing whether a result is statistically significant or not, Chopra et al. (2024) randomize whether the studies include an expert prediction of the effect size and

whether the statistical uncertainty is communicated in terms of p-values or standard errors. This allows for richer insights into whether the null result penalty depends on whether a result is considered "surprising" by experts and whether focusing on p-values amplifies the bias. Chopra et al. (2024) also use a number of "obfuscation treatments" to further mitigate concerns about experimenter demand effects and SDB. Specifically, they cross-randomize the seniority of the research team and the ranking of their universities. When including such obfuscation treatments, it is especially important to submit a pre-analysis plan that clarifies the main comparisons of interest.

One limitation of vignettes arises from the potentially low interpersonal comparability of responses given heterogeneous interpretation of response options. To improve interpersonal comparability, King et al. (2004) develop anchoring vignettes. These vignettes try to correct for differential interpretation of response categories across individuals by using standardized hypothetical scenarios.

4.7 Incentives

Incentives are a standard tool in experimental economics, where researchers routinely tie payoffs to verifiable outcomes to elicit truthful responses and minimize noise. The key idea is that incentives make misreporting financially costly and thereby increase the truthfulness of responses. However, this approach is often not feasible in the context of self-reported data, which typically lack an objective benchmark for correctness (e.g., Harrison and Swarthout, 2025).[11] As a result, many survey-based studies must rely on unincentivized responses, leaving them vulnerable to various forms of bias (e.g., Harrison, 2006, 2024)—including SDB.

Procedure When belief elicitation involves an objective external benchmark—such as a factual economic indicator—respondents can be rewarded for stating more accurate beliefs. For discrete outcomes, this may involve a reward for correct answers; for continuous outcomes, incentives can be tied to proximity within a specified range (Hossain and Okui, 2013; Schotter and Trevino, 2014). These mechanisms are intuitive, simple to implement, and create clear stakes for truthful reporting. However, they are only incentive-compatible for eliciting the mode of a respondent's

[11] When objective benchmarks are unavailable, as is often the case in belief elicitation, the Bayesian Truth Serum offers a workaround by aligning incentives for truthful reporting by rewarding answers that are more common than predicted by others (Prelec, 2004).

belief distribution (cf. Harrison and Swarthout, 2025).

Assumptions and Pitfalls Using incentives to reduce SDB rests on several assumptions. First, it assumes respondents understand the incentive mechanism and trust that payouts are contingent on accuracy rather than socially desirable responses. Second, incentives presume that respondents are motivated primarily by monetary gains; this may not hold in high-stakes political or identity-laden contexts where expressive utility might dominate. Hence, while incentives can be powerful, they are not a universal fix and must be used with careful design and contextual awareness.

Best Practices Effective use of incentives requires clarity and restraint. Incentive schemes should be simple and easy to understand, as complex designs can, in some instances, reduce truthfulness by creating confusion (Danz et al., 2022). Pre-testing mechanisms in the survey population is essential to catch misunderstandings and ensure the incentives function as intended.

To draw correct conclusions, it is advisable to rely on rich data: By eliciting not only modes of beliefs but instead assessing full belief distributions, measuring different objects (such as higher-order beliefs), or including repeated measurement, hidden patterns in the data can be uncovered (e.g., Harrison and Swarthout, 2025; Harrison et al., 2025, 2017; Hartzmark and Sussman, 2024). Specifically, confidence in beliefs can be extracted this way (e.g., Harrison and Swarthout, 2022). In a similar vein, the choice of the underlying event—binary, continuous, or categorical—poses an important design choice that influences the effectiveness of measurement, and hence should be chosen carefully.

Evidence on Effectiveness Incentives can help reduce SDB by shifting respondents' motivation from self-presentation to accurate reporting. In contexts where beliefs may be motivated or expressive—such as in politics—accuracy incentives have been shown to reduce partisan bias. For instance, prediction incentives narrow the partisan gap in beliefs about unemployment (e.g., Bullock et al., 2015; Prior et al., 2015). They also modestly reduce political polarization (Peterson and Iyengar, 2021), though effects are weaker for rumors (Berinsky, 2018) and absent in some domains, such as COVID-19 beliefs (Allcott et al., 2020). Zimmermann (2020) shows that incentives strongly decrease the extent of motivated forgetting in the context of image-relevant information. Overall, while incentives can reduce SDB—particularly for politically motivated

Method	Rationale	Procedure	Key Assumptions	Potential Pitfalls	Evidence on Effectiveness
List Experiment	Mask individual answers in aggregates	Compare number of endorsed items in treatment vs. control lists	No design effects; truthful inclusion of sensitive item	Miscounting; high variance; sensitive item may still draw attention	Yields higher reporting of sensitive behaviors than direct questions; large heterogeneity (Gilligan et al., 2024; Li and Van den Noortgate, 2022); some evidence for confusion and false positives (Markhof et al., 2025)
Randomized Response Technique	Ensure plausible deniability	Use private randomization device to determine response	Respondents follow protocol and trust privacy	Confusion; strategic misuse; high variance	Significant scope for bias (John et al., 2018)
Anonymity Guarantees	Reduce perceived material and social image costs	Emphasize privacy; use self-administered surveys	Trust in anonymity guarantees	May reduce engagement or raise suspicion	Boosts truthfulness; self-administered surveys outperform interviews (Kreuter et al., 2008)
Third-Person / Social-Circle	Reduce self-image concerns by shifting perspective	Ask about peers' behavior instead of own	Respondents project own attitudes onto peers	Pluralistic ignorance; weak correlation in heterogeneous groups	Often reveals higher prevalence of sensitive traits (Fisher, 1993)
Forgiving Outcome Framing	Use normalizing language that reduces stigma	Frame the question in more forgiving terms	Alters the extent of perceived stigma	Could backfire by making the sensitivity more salient	Mixed evidence on effectiveness (Kaplan and Yu, 2015; Näher and Krumpal, 2012; Peter and Valkenburg, 2011).
Vignette Experiments	Sensitive issue less explicit	Formulate hypothetical scenario with many characteristics	Answers to hypothetical scenario are externally valid	Long vignettes lead to decreased attention; "hiding" sensitive issue might not be successful	Limited evidence
Incentives	Shift motivation from self-presentation to accuracy	Provide monetary rewards tied to accuracy against an objective benchmark	Respondents understand and trust incentives	Risk of strategic guessing or searching; limited to beliefs with objective benchmarks; misunderstanding incentives	Effective in reducing partisan bias in factual beliefs (Bullock et al., 2015; Prior et al., 2015)

Table 1: Overview of Methods for Detecting or Mitigating Social Desirability Bias (SDB)

beliefs—their effectiveness depends on context, the availability of information, and the nature of the belief being elicited.

5 Matching Design Tools to Context

The effectiveness of any strategy to mitigate SDB depends crucially on understanding the underlying motivation behind misreporting. While often grouped under a single label, SDB can arise from three distinct sources: material costs, social-image concerns and self-image concerns. In addition, a method's effectiveness hinges on its complexity, that impacts respondent understanding and potential confusion.

First, consider the case where respondents may fear direct material costs. This case is particularly relevant in surveys on illegal behavior, politically sensitive attitudes, or stigmatized health conditions. In these contexts, respondents may worry that their answers could be traced back to them, even if anonymous, or that participation itself carries risk. Such settings thus require techniques that most credibly provide anonymity. Traditionally, the RRT is recommended in such settings as the probabilistic element creates a credible cover for responses. However, its effectiveness critically hinges on trust and high levels of understanding. Indeed, the RRT is probably the most complex method used to mitigate SDB and might therefore induce non-strategic reporting errors causing false positives (Markhof et al., 2025). A less complex alternative to the RRT which similarly provides a credible cover is the LE, which may be the method of choice for populations less familiar with probabilistic survey items.

When social-image concerns dominate, respondents are motivated by reputational considerations. They may worry about appearing immoral, prejudiced, irresponsible, or otherwise norm-violating in the eyes of others, including researchers or enumerators. Just like for material cost, the most effective methods are privacy enhancing methods. LEs and strong anonymity guarantees all serve this purpose by providing a cover. These techniques can meaningfully reduce the perceived social cost of disclosure, but are less complex than RRTs. Yet, LEs are somewhat more complex than direct question methods.

Self-image concerns arise when individuals feel psychological discomfort in acknowledging behaviors or attitudes that conflict with their internal moral standards or ideal self-image. Even in the absence of observers or external judgment, respondents may misreport to maintain a favorable view of themselves. In such cases, privacy-enhancing methods like LEs and RRT are ineffective, since they do not necessarily shield the research objective from the respondent. Instead, other approaches—such as third-person framing or forgiving outcome framing—can help reduce psychological resistance by distancing the respondent from the sensitive content.

Incentives are an alternative tool as they raise motivation to answer truthfully despite of the psychological cost of truthful responding.

Table 2 summarizes the most commonly used tools for mitigating SDB and offers guidance on when each is most appropriately applied. Ultimately, mitigating SDB requires an understanding of what the respondent is trying to protect: their self-image, their social image, or their material well-being. Finally, it is essential to consider the complexity of the method, particularly when dealing with populations less familiar with probabilistic reasoning, as they might find more sophisticated techniques, like the RRT, confusing.

Method	Best Applied When
Randomized Response	Questions involve highly sensitive or legally risky behaviors and respondents understand complex instructions.
List Experiment	Respondents may misreport due to fear of judgment when answering sensitive questions; researcher wants to avoid too complex instructions (as with RRT).
Anonymity Guarantees	Respondents misreport due to social image concerns. Researcher surveys respondents that struggle with more complex instructions (as in LE and RRT).
Third-Person / Social Circle	Respondents may distort self-reports due to self-image concerns; shifting focus to others allows inference through projection.
Forgiving Outcome Framing	Self-image concerns distort responses, but no accuracy incentives are feasible.
Vignette Experiments	Educated population which understands hypothetical scenarios so that external validity is plausible; vignette can be designed such that the sensitive issue is not too salient.
Incentives	Truthful responses are psychologically costly due to self or social image concerns and credible benchmarks for incentives are available; Respondents understand complex incentives.

Table 2: Toolkit for Mitigating Social Desirability Bias

6 When Social Desirability Bias Distorts Treatment Effects

In some experiments, the treatment not only affects the outcome of interest but also plausibly shifts what respondents perceive as the socially desirable answer. For instance, when researchers provide information suggesting that immigration does not harm labor market outcomes, the treatment may influence respondents' stated policy preferences not just by changing their beliefs, but by altering what they feel they ought to say. In such cases, SDB may differentially affect responses in the treatment and control groups, distorting the estimated treatment effect.[12]

[12] See de Quidt et al. (2018) for a formal discussion of estimating treatment effects in the presence of experimenter demand effects.

Research designs that aim to identify the persuasive effect of information must therefore attempt to isolate—and, where feasible, purge—this reporting channel.

To deal with this issue, we outline a series of best practices for minimizing the impact of SDB on treatment effect estimation in experimental settings, building on Haaland et al. (2023). It is important to note that some of the techniques discussed in Section 4—such as LEs—introduce substantial noise and therefore require very large sample sizes to maintain adequate statistical power. Therefore, these methods are less suited when estimating treatment effects and are not covered in this subsequent section. Table 3 provides an overview of different methods that are specifically tailored towards mitigating distortions of SDB on treatment effects, which we discuss below.

Table 3: Toolkit for Mitigating Social Desirability Bias in Treatment Effect Estimation

Method	Description	Use Case	Limitations
Natural Field Experiments	Participants are unaware they are part of an experiment; behavior occurs in natural contexts.	Useful when outcomes can be observed passively.	High cost, difficult to implement, limited to observable behaviors.
Anonymity	Use group-level or anonymized individual outcomes (e.g., anonymous petitions).	Suitable for sensitive topics (e.g., politics, prejudice).	Limits ability to link outcomes to individual characteristics.
Obfuscated Follow-Ups	Recontact participants using independent cover story to separate outcome measurement from treatment.	Effective for self-reported outcomes.	Attrition; some residual linkage may persist.
Purpose Obfuscation	Introduce a cover story to obscure hypothesis.	Best when treatment content may potentially reveal normative expectations.	Can dilute treatment intensity or raise ethical concerns.
Measuring Perceived Norms	Ask respondents directly what they believe is socially desirable.	Enables identification of differential SDB across treatments.	Self-reports may not capture true underlying beliefs.
Heterogeneity by Norm Sensitivity	Use psychometric scales to measure individual differences in SDB susceptibility.	Helps identify whether SDB moderates treatment effects.	Heterogeneity tests are often hard to interpret.

Natural Field Experiments A powerful tool for mitigating SDB is the use of natural field experiments, in which participants are unaware that they are part of an experiment and behave as they would in their everyday lives (Harrison and List, 2004). By embedding experimental interventions in real-world contexts with naturally occurring behaviors, researchers can often observe outcomes that are less contaminated by social desirability concerns—both because there

is no salient experimenter to trigger social image motivations, and because the stakes are typically higher, increasing the cost of misrepresentation.

Anonymity One solution is to use outcomes only observed at the group level and not at the individual level. For example, Grigorieff et al. (2020) use anonymous online petitions to measure private political opinions, where the researcher cannot tell if an individual has signed the petition or not, but instead only observes the number of people who signed the petition in different treatment groups respectively.

Obfuscated Follow-Ups Haaland and Roth (2020) propose obfuscated follow-ups as a strategy to mitigate differential SDB across treatment conditions. These follow-up studies recontact the original participants but are presented as separate and unrelated, thereby severing the connection between treatment assignment and outcome measurement. This makes differential SDB between treatment and control much less likely. For instance, if a respondent feels social pressure to support higher immigration immediately after being informed that there are no adverse labor market impacts of immigration, the differential social pressure induced by the treatment should disappear in a seemingly unrelated study in which the researcher is perceived to be unaware of the earlier information provision (Haaland and Roth, 2020).

Purpose Obfuscation One way to mitigate the possibility that there is differential desirability bias across treatment conditions is to obscure the cues that indicate which responses are socially desirable. For example, Bursztyn et al. (2020b) implement an experimental design that avoids linking the researchers with any specific political ideology: Participants are informed they will have the option to donate to a randomly drawn organization—either anti-immigration or pro-immigration.

Another approach is to introduce a cover story that redirects attention away from the normative implications of the treatment, making it more likely that social desirability pressures are similar across conditions. For instance, in an information provision experiment on the share of the immigrant population, Hopkins et al. (2019) use the following cover story when administering their treatment: "We are interested in whether you've heard about a story that has been in the news. The story is: ..." By framing the treatment as a news awareness question rather than as a corrective message that should shape one's views, the researchers reduce the risk that the treatment gives additional cues about which response is socially preferred. More generally, when

participants are unsure which answers the researcher considers socially desirable, differences in SDB across treatment arms are less likely to confound results (de Quidt et al., 2018).

Measuring What Is Perceived as Socially Desirable. A direct approach to assessing differential SDB across treatment groups is to measure it explicitly. One option is to elicit participants' beliefs about what constitutes socially desirable behavior in the survey using an open-ended question, ideally at the very end. This sheds light on whether perceptions of what is socially desirable may vary systematically by treatment. These open-ended data are naturally limited by their self-reported nature (Haaland et al., 2025) and are subject to SDB themselves. An option that partly circumvents this issue is to measure several objects related to the social norm of interest, being first-order normative, first-order descriptive, second-order normative, and second-order descriptive beliefs (Harrison and Swarthout, 2025). This approach allows to evaluate consistency between incentivized and hypothetical measures (e.g., Harrison, 2006, 2024).

Heterogeneity by Sensitivity to Social Norms. If SDB influences experimental responses, its effects should be most pronounced among individuals who are both attuned to social expectations and willing to adjust their behavior accordingly. This can be tested by examining whether treatment effects vary with individual differences in sensitivity to social norms, measured by scores on established psychometric scales (Allcott and Taubinsky, 2015; Dhar et al., 2022). Commonly used measures include the Marlowe–Crowne Social Desirability Scale (Crowne and Marlowe, 1960), abbreviated versions of this scale (e.g., Nießen et al., 2019), and the Self-Monitoring Scale (Snyder, 1974). These scales capture different aspects of social conformity and impression management, ranging from internal dispositions toward self-enhancement to outward-oriented concerns with social approval. However, as noted by de Quidt et al. (2025), these heterogeneity tests are often hard to interpret and inconclusive as the traits measured in the scales may be correlated with underlying preferences, which may be responsible for a different response to the treatment.

7 Conclusions

SDB remains a pervasive challenge in survey-based and experimental research, systematically distorting self-reports in ways that obscure true attitudes and behaviors. While decades of work across disciplines have produced a rich set of tools to detect and mitigate SDB—from indirect questioning formats like LEs, forgiving outcome framing, anonymity guarantees and

incentives—no single method universally resolves the issue. Each approach comes with tradeoffs, often requiring researchers to balance potential reductions in bias against increases in statistical noise or complexity.

This review highlights several key takeaways. First, the prevalence and magnitude of SDB vary significantly across domains, shaped by the sensitivity of the topic. Second, many commonly used strategies to mitigate SDB sometimes fail by creating confusion or inadvertently increasing perceived sensitivity. Third, SDB can not only affect measurement of levels but can also distort estimated treatment effects when treatments shift perceived norms.

Going forward, we encourage researchers to approach SDB as a context-dependent distortion that can often be anticipated and measured. Mitigation strategies should be selected based on the dominant mechanism behind the bias—whether it stems from self-image, social-image, or fear of material costs—and matched to the specific goals and constraints of the research design.

A key priority for future research is the development of more large-scale validation studies that combine individual-level ground-truth data with alternative elicitation techniques. Such studies would allow researchers to directly compare the accuracy of different methods. Equally important is understanding how SDB operates in increasingly digital survey environments. As researchers experiment with video interviews and AI-led qualitative interviews (Chopra and Haaland, 2023), it remains unclear whether these new modes attenuate or amplify social desirability pressures.

References

Ajzenman, Nicolas, Guillermo Cruces, Ricardo Perez-Truglia, Darío Tortarolo, and Gonzalo Vazquez-Bare, "From Flat to Fair? The Effects of a Progressive Tax Reform," Technical Report, National Bureau of Economic Research 2024.

Alem, Yonas, Håkan Eggert, Martin G Kocher, and Remidius D Ruhinduka, "Why (field) experiments on unethical behavior are important: Comparing stated and revealed behavior," *Journal of Economic Behavior & Organization*, 2018, *156*, 71–85.

Alexander, Cheryl S. and Henry Jay Becker, "The Use of Vignettes in Survey Research," *The Public Opinion Quarterly*, 1978, *42* (1).

Allcott, Hunt and Dmitry Taubinsky, "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market," *American Economic Review*, 2015, *105* (8), 2501–2538.

_ , Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Y. Yang, "Polarization and Public Health: Partisan Differences in Social Distancing during the Coronavirus Pandemic," *Journal of Public Economics*, 2020, *191*, 104254.

Archambault, Patrick M., Rhonda J. Rosychuk, Martyne Audet, Rajan Bola, Shahram Vatankour, Steven C. Brooks, Roual Daoust, Gregory Clark, Lara Grant, Samuel Vaillancourt, Michelle Welsford, Laurie J. Morrison, and Corinne M. Hohl, "Accuracy of

Self-Reported COVID-19 Vaccination Status Compared With a Public Health Vaccination Registry in Québec: Observational Diagnostic Study," *JMIR Public Health and Surveillance*, 2023, *9* (1), e44465.

Barmettler, Franziska, Ernst Fehr, and Christian Zehnder, "Big Experimenter is Watching You! Anonymity and Prosocial Behavior in the Laboratory," *Games and Economic Behavior*, 2012, *75*, 17–34.

Bekkers, René and Pamala Wiepking, "Accuracy of self-reports on donations to charitable organizations," *Quality & Quantity*, 2011, *45*, 1369–1383.

Belli, Robert F., Sean E. Moore, and John VanHoewyk, "An Experimental Comparison of Question Forms Used to Reduce Vote Overreporting," *Electoral Studies*, 2006, *25* (4), 751–759.

Bénabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

_ and _, "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 2016, *30* (3), 141–164.

Berinsky, Adam J., "Telling the Truth about Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding," *The Journal of Politics*, 2018, *80* (1), 211–224.

Bertrand, Marianne and Sendhil Mullainathan, "Do People Mean What They Say? Implications for Subjective Survey Data," *American Economic Review*, 2001, *91* (2), 67–72.

Bicchieri, Cristina, *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press, 2006.

Blair, Graeme, Alexander Coppock, and Margaret Moor, "When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments," *American Political Science Review*, 2020, *114* (4), 1297–1315.

_ and Kosuke Imai, "Statistical analysis of list experiments," *Political Analysis*, 2012, *20* (1), 47–77.

_ , _ , and Yang-Yang Zhou, "Design and Analysis of the Randomized Response Technique," *Journal of the American Statistical Association*, 2015, *110* (511), 1304–1319.

Boudreau, Laura E., Sylvain Chassang, Ada Gonzalez-Torres, and Rachel Heath, "Monitoring Harassment in Organizations," NBER Working Paper No. 31011, National Bureau of Economic Research 2023.

Bradburn, Norman, "Respondent Burden," in "Proceedings of the Survey Research Methods Section of the American Statistical Association," Vol. 35 American Statistical Association Alexandria, VA, USA 1978, pp. 35–40.

Braghieri, Luca, "Political Correctness, Social Image, and Information Transmission," *American Economic Review*, 2024, *114* (12), 3877–3904.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber, "Partisan Bias in Factual Beliefs about Politics," *Quarterly Journal of Political Science*, 2015, *10* (4), 519–578.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, 2020, *110* (10), 2997–3029.

_ , Alexander W. Cappelen, Bertil Tungodden, Alessandra Voena, and David H. Yanagizawa-Drott, "How Are Gender Norms Perceived?," NBER Working Paper No. 31049, National Bureau of Economic Research 2023.

_ and Robert Jensen, "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure," *Annual Review of Economics*, 2017, *9*, 131–153.

_ , **Georgy Egorov, and Stefano Fiorin**, "From Extreme to Mainstream: The Erosion of Social Norms," *American Economic Review*, 2020, *110* (11), 3522–3548.

_ , _ , **Ingar Haaland, Aakaash Rao, and Christopher Roth**, "Justifying Dissent," *Quarterly Journal of Economics*, 2023, *138* (3), 1403–1451.

Catania, Joseph A., Diane Binson, Jesse Canchola, Lance M. Pollack, Walter Hauck, and Thomas J. Coates, "Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior," *Public Opinion Quarterly*, 1996, *60* (3), 345–375.

Chakravarty, Anujit, Arkadev Ghosh, Matt Lowe, and Gareth Nellies, "Learning About Outgroups: The Impact of Broad Versus Deep Interactions," 2022. CESifo Working Paper No. 11363.

Chan, Wei-Hung, Ching-Huang Lai, Shu-Jia Huang, Chung-Chi Huang, Chung-Yu Lai, Yi-Chun Liu, Shiang-Huei Jiang, Shan-Ru Li, Ya-Mei Tzeng, Senyeong Kao, Yu-Tien Chang, Chia-Chao Wu, Chao-Yin Kuo, Kuang-Chen Hung, and Yu-Lung Chiu, "Verifying the accuracy of self-reported smoking behavior in female volunteer soldiers," *Scientific Reports*, 2023, *13* (3438).

Chen, Yuyu and David Y. Yang, "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review*, 2019, *109* (6), 2294–2332.

Chopra, Felix and Ingar Haaland, "Conducting qualitative interviews with AI," 2023.

_ , _ , **Christopher Roth, and Andreas Stegmann**, "The null result penalty," *The Economic Journal*, 2024, *134* (657), 193–219.

Chuang, Erica, Pascaline Dupas, Elise Huillery, and Juliette Seban, "Sex, Lies, and Measurement: Consistency Tests for Indirect Response Survey Methods," *Journal of Development Economics*, 2021, *148*, 102582.

Coffman, Katherine B., Lucas C. Coffman, and Keith M. Marzilli Ericson, "The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated," *Management Science*, 2017, *63* (10), 3147–3529.

Cohn, Alain, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd, "Civic honesty around the globe," *Science*, 2019, *365* (6448), 70–73.

Colley, Rachel C, Gregory Butler, Didier Garriguet, Stephanie A Prince, and Karen C Roberts, "Comparison of self-reported and accelerometer-measured physical activity in Canadian adults," *Health Reports*, 2018, *29* (12), 3–15.

Cortés, Patricia, Gizem Koşar, Jessica Pan, and Basit Zafar, "Should Mothers Work? How Perceptions of the Social Norm Affect Individual Attitudes toward Work in the US," *Review of Economics and Statistics*, 2024, pp. 1–28.

Coutts, Elisabeth and Ben Jann, "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)," *Sociological Methods & Research*, 2011, *40* (1), 169–193.

Crowne, Douglas P. and David Marlowe, "A New Scale of Social Desirability Independent of Psychopathology," *Journal of Consulting Psychology*, 1960, *24* (4), 349–354.

Danz, David, Lise Vesterlund, and Alistair J. Wilson, "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 2022, *112* (9), 2851–2883.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

_ , **Lise Vesterlund, and Alistair J. Wilson**, "Experimenter Demand Effects," in Alex Rees-Jones, ed., *Handbook of Experimental Methods in the Social Sciences*, Edward Elgar Publishing, 2025.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao, "Voting to Tell Others," *The Review of Economic Studies*, 2017, *84* (1), 143–181.

Dhar, Diva, Tarun Jain, and Seema Jayachandran, "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India," *American Economic Review*, 2022, *112* (3), 899–927.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman, "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism," *American Economic Review*, 2015, *105* (11), 3416–3442.

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk, "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias," Working Paper 29549, National Bureau of Economic Research December 2021.

Edwards, Allen L., *The Social Desirability Variable in Personality Assessment and Research*, New York: Dryden Press, 1957.

Ehler, Ingmar, Felix Wolter, and Justus Junkermann, "Sensitive Questions in Surveys: A Comprehensive Meta-Analysis of Experimental Survey Studies on the Performance of the Item Count Technique," *Public Opinion Quarterly*, 2021, *85* (1), 6–27.

Enke, Benjamin and Thomas Graeber, "Cognitive Uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Ewers, Mara and Florian Zimmermann, "Image and Misreporting," *Journal of the European Economic Association*, 2015, *13* (2), 363–380.

Farrell, Joseph and Matthew Rabin, "Cheap talk," *Journal of Economic perspectives*, 1996, *10* (3), 103–118.

Fisher, Robert J., "Social Desirability Bias and the Validity of Indirect Questioning," *Journal of Consumer Research*, 1993, *20* (2), 303–315.

Fuster, Andreas and Basit Zafar, "Chapter 4 - Survey Experiments on Economic Expectations," in Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, eds., *Handbook of Economic Expectations*, Elsevier, 2023, pp. 107–130.

Galesic, Mirta, Wändi Bruine de Bruin, Mariel Dumas, Arie Kapteyn, Jonathan E. Darling, and Erik Meijer, "Asking about Social Circles Improves Election Predictions," *Nature Human Behaviour*, 2018, *2*, 187–193.

Gilligan, Daniel O., Melissa Hidrobo, Jessica Leight, and Heleene Tambet, "Using a List Experiment to Measure Intimate Partner Violence: Cautionary Evidence from Ethiopia," *Applied Economics Letters*, 2024, pp. 1–7.

Glynn, Adam N, "What can we learn with statistical truth serum? Design and analysis of the list experiment," *Public Opinion Quarterly*, 2013, *77* (S1), 159–172.

Grigorieff, Alexis, Christopher Roth, and Diego Ubfal, "Does Information Change Attitudes toward Immigrants?," *Demography*, 2020, *57*, 1117–1143.

Haaland, Ingar and Christopher Roth, "Labor Market Concerns and Support for Immigration," *Journal of Public Economics*, 2020, *191*, 104256.

_ , _ , and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2023, *61* (1), 3–40.

Haaland, Ingar K, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-Ended Survey Data," *Journal of Economic Literature*, 2025.

Hainmueller, Jens and Dominik Hangartner, "Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination," *American Political Science Review*, 2013, *107* (1), 159–187.

Haire, Mason, "Projective Techniques in Marketing Research," *Journal of Marketing*, 1950, *14* (5), 649–656.

Harrison, Glenn W., "Hypothetical Bias over Uncertain Outcomes," in John A. List, ed., *Using Experimental Methods in Environmental and Resource Economics*, Edward Elgar Publishing, 2006.

_ , "Real Choices and Hypothetical Choices," in Stephane Hess and Andrew Daly, eds., *Handbook of Choice Modelling*, Edward Elgar Publishing, 2024, pp. 246–275.

_ **and J. Todd Swarthout**, "Belief Distributions, Bayes Rule and Bayesian Overconfidence," CEAR Working Paper 2020-11, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University 2022.

_ **and** _ , "Causal Inferences Over Unobservables," CEAR Working Paper 2025-02, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University 2025.

_ **and John A. List**, "Field Experiments," *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.

_ **, Don Ross, and J. Todd Swarthout**, "Gender, Confidence, and the Mismeasure of Intelligence, Competitiveness and Literacy," *Journal of Political Economy*, 2025.

_ **, Jimmy Martínez-Correa, J. Todd Swarthout, and Eric R. Ulm**, "Scoring Rules for Subjective Probability Distributions," *Journal of Economic Behavior & Organization*, 2017, *134*, 430–448.

Hartzmark, Samuel M. and Abigail B. Sussman, "Eliciting Expectations," Working Paper, Boston College Carroll School of Management and NBER 2024.

Henkel, Luca, Roland Bénabou, Armin Falk, and Jean Tirole, "Eliciting Moral Preferences under Image Concerns: Theory and Experiment," 2024.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review*, 1996, *86* (3), 653–660.

Höglinger, Marc and Ben Jann, "More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model," *PloS one*, 2018, *13* (8), e0201770.

Holbrook, A. L. and J. A. Krosnick, "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique," *Public Opinion Quarterly*, 2010, *74* (1), 37–67.

Hopkins, Daniel J., John Sides, and Jack Citrin, "The Muted Consequences of Correct Information about Immigration," *The Journal of Politics*, 2019, *81* (1).

Hossain, Tanjim and Ryo Okui, "The binarized scoring rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Höglinger, Marc and Andreas Diekmann, "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT," *Political Analysis*, 2017, *25*, 131–137.

Jiao, Peiran, "Payoff-Based Belief Distortion," *The Economic Journal*, 2020, *130* (629), 1416–1444.

John, Leslie K., George Loewenstein, Alessandro Acquisti, and Joachim Vosgerau, "When and Why Randomized Response Techniques (Fail to) Elicit the Truth," *Organizational Behavior and Human Decision Processes*, 2018, *148*, 101–123.

Kaplan, Robin L. and Erica C. Yu, "Measuring Question Sensitivity," *American Association for Public Opinion Research*, 2015, pp. 4107–4121.

Karp, Jeffrey A and David Brockington, "Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries," *The Journal of Politics*, 2005, *67* (3), 825–840.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon, "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research," *American Political Science Review*, 2004, *98* (1), 191–207.

Kirchner, Antje, "Validating Sensitive Questions: A Comparison of Survey and Register Data," *Journal of Official Statistics*, 2015, *31* (1), 31–59.

Kleven, Øyvin, "The effect of nonresponse and measurement errors on turnout estimates: Electronic publication.

_ **and Kristen Ringdal**, "Causes and effects of measurement errors in educational attainment: Experiences from The European Social Survey in Norway," Documents 2020/35, Statistics Norway 2020.

_ **and Tove Bergseteren**, "Ulike innsamlingsmåter i Velgerundersøkelsen 2019: Effekter av å benytte telefonintervju eller webintervju," Notater / Documents 2023/34, Statistisk sentralbyrå, Oslo June 2023. Elektronisk publikasjon.

Kraay, Aart and Peter Murrell, "Misunderestimating corruption," *Review of Economics and Statistics*, 2016, *98* (3), 455–466.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau, "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity," *Public Opinion Quarterly*, 2008, *72* (5), 847–865.

Kuhn, Patrick M. and Nick Vivyan, "The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries," *Political Analysis*, 2022, *30* (2).

Kuklinski, James H., Michael D. Cobb, and Martin Gilens, "Racial Attitudes and the "New South"," *The Journal of Politics*, 1997, *59* (2), 323–349.

Kuran, Timur, *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Harvard University Press, 1998.

Kvalvik, Liv G., Roy M. Nilsen, Rolv Skjærven, Stein Emil Vollset, Øivind Midttun, Per Magne Ueland, and Kjell Haug, "Self-reported smoking status and plasma cotinine concentrations among pregnant women in the Norwegian Mother and Child Cohort Study," *Pediatric Research*, 2012, *72* (1), 101–107.

Lamb, Charles W. and Donald E. Stem, "An Empirical Validation of the Randomized Response Technique," *Journal of Marketing Research*, November 1978, *15* (4), 616–621.

Lease, Matthew, Jessica Hullman, Jeffrey Bigham, Michael Bernstein, Juho Kim, Walter Lasecki, Saeideh Bakhshi, Tanushree Mitra, and Robert Miller, "Mechanical Turk is Not Anonymous," 2013. Available at SSRN: https://ssrn.com/abstract=2228728.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas, "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation," *Sociological Methods & Research*, 2005, *33* (3), 319–348.

Levitt, Steven D. and John A. List, "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

Li, Jiayuan and Wim Van den Noortgate, "A Meta-Analysis of the Relative Effectiveness of the Item Count Technique Compared to Direct Questioning," *Sociological Methods & Research*, 2022, *51* (2), 760–799.

Ling, Yier and Alex Imas, "Underreporting of AI use: The role of social desirability bias," *Available at SSRN*, 2025.

List, John A., "Valuing Non-Marketed Goods and Services Using a List Experiment in the Field," 2025. Available at SSRN: `https://ssrn.com/abstract=5096019`.

Locander, William, Seymour Sudman, and Norman Bradburn, "An Investigation of Interview Method, Threat and Response Distortion," *Journal of the American Statistical Association*, June 1976, *71* (354), 269–275.

Loewenstein, George, "Experimental Economics from the Vantage-Point of Behavioural Economics," *The Economic Journal*, 1999, *109* (453), 25–34.

Markhof, Yannick, Stephan Dietrich, and Rose Camille Vincent, "Lies in Disguise: Measurement Error in Popular Survey Methods for Sensitive Issues," 2025. Draft.

Miller, Judith Droitcour, *A New Survey Technique for Studying Deviant Behavior*, The George Washington University, 1984.

Näher, Anatol-Fiete and Ivar Krumpal, "Asking Sensitive Questions: The Impact of Forgiving Wording and Question Context on Social Desirability Bias," *Quality & Quantity*, 2012, *46*, 1601–1616.

Nießen, Désirée, Melanie V Partsch, Christoph J Kemper, and Beatrice Rammstedt, "An English-language adaptation of the social desirability–gamma short scale (KSE-G)," *Measurement Instruments for the Social Sciences*, 2019, *1*, 2–2019.

Nyqvist, Martina Björkman, Lucia Corno, Damien De Walque, and Jakob Svensson, "Incentivizing safer sexual behavior: evidence from a lottery experiment on HIV prevention," *American Economic Journal: Applied Economics*, 2018, *10* (3), 287–314.

Paulhus, Delroy L., "Two-Component Models of Socially Desirable Responding," *Journal of Personality and Social Psychology*, 1984, *46* (3), 598–609.

Peter, Jochen and Patti M. Valkenburg, "The Impact of "Forgiving" Introductions on the Reporting of Sensitive Behavior in Surveys: The Role of Social Desirability Response Style and Developmental Status," *Public Opinion Quarterly*, 2011, *75* (4), 779–787.

Peterson, Erik and Shanto Iyengar, "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?," *American Journal of Political Science*, 2021, *65* (1), 133–147.

Prelec, Drazen, "A Bayesian Truth Serum for Subjective Data," *Science*, 2004, *306* (5695), 462–466.

Prior, Markus, Gaurav Sood, and Kabir Khanna, "You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions," *Quarterly Journal of Political Science*, 2015, *10* (4), 489–518.

Reisinger, James, "Subjective Well-Being and Social Desirability," *Journal of Public Economics*, 2022, *214*, 104745.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro, "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions," *American Journal of Political Science*, 2016, *60* (3), 783–802.

Schotter, Andrew and Isabel Trevino, "Belief Elicitation in the Laboratory," *Annual Review of Economics*, 2014, *6* (1), 103–128.

Snyder, Mark, "Self-Monitoring of Expressive Behavior," *Journal of Personality and Social Psychology*, 1974, *30* (4), 526–537.

Stantcheva, Stefanie, "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible," *Annual Review of Economics*, 2023, *15*, 205–234.

Sudman, Seymour and Norman M. Bradburn, *Response Effects in Surveys: A Review and Synthesis*, Chicago: Aldine Publishing, 1974.

Sweeney, Latanya, "Simple Demographics Often Identify People Uniquely," Working Paper 3, Carnegie Mellon University 2000.

Tourangeau, Roger and Ting Yan, "Sensitive Questions in Surveys," *Psychological Bulletin*, 2007, *133* (5), 859–883.

Tracy, Paul E. and James Alan Fox, "The Validity of Randomized Response for Sensitive Measurements," *American Sociological Review*, April 1981, *46* (2).

van der Heijden, Peter G. M., Ger van Gils, Jan Bouts, and Joop J. Hox, "A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning," *Sociological Methods & Research*, May 2000, *28* (4).

Warner, Stanley L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 1965, *60* (309), 63–69.

Wolter, Felix and Peter Preisendörfer, "Asking Sensitive Questions: An Evaluation of the Randomized Response Technique vs. Direct Questioning Using Individual Validation Data," *Sociological Methods Research*, 2013, *42*.

Yan, Ting, "Consequences of Asking Sensitive Questions in Surveys," *Annual Review of Statistics and Its Application*, 2021, *8*, 109–127.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang, "Two new models for survey sampling with sensitive characteristic: design and analysis," *Metrika*, 2008, *67*, 251–263.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2020, *110* (2), 337–361.

Online Appendix

The Social Desirability Atlas

Leonardo Bursztyn Ingar Haaland
Christopher Roth Nicolas Röver

May 26, 2025

A Comparative Validation Studies with Individual Level Benchmarks

Appendix Table A1: Summary of Studies Using Verifiable Individual-Level Benchmarks to Evaluate SDB and Mitigating Techniques

Reference	Outcome	Sample Size	True Prevalence	Technique	Estimate	95% CI	Accuracy
Höglinger and Jann (2018)	Cheating in Online Experiment 1 (Roll-a-six Game)	382	4%	Direct Question	4%	[0.02; 0.06]	98%
		1145	6%	RRT Variation 1	14%	[0.1; 0.18]	86%
		780	5%	RRT Variation 2	5%	[0.02; 0.08]	95%
		771	5%	RRT Variation 3	–2%	[–0.05; 0.01]	97%
	Cheating in Online Experiment 2 (Prediction Game)[a]	387	24%	Direct Question	2%	[–0.13; 0.17]	79%
		1168	27%	RRT Variation 1	15%	[0.11; 0.19]	73%
		760	26%	RRT Variation 2	4%	[0.01; 0.07]	78%
		759	27%	RRT Variation 3	1%	[–0.03; 0.05]	76%
John et al. (2018)	Cheating when Self-Grading a Test	66	61%	Direct Question	27%	–[b]	-[b]
		132	55%	RRT	3%	–[b]	–[b]
	Lying about Current Location (Study 3)	151	23%[c]	Direct Question	19%	–[b]	–[b]
		605	23%[c]	RRT	–26%	–[b]	–[b]
Kirchner (2015)	Receipt of Unemployment Benefits Among Unemployed	579	1	Direct Question	87%	[0.84; 0.9]	87%
		836		RRT	85%	[0.81; 0.89]	85%
Kuhn and Vivyan (2022)	Non-Voting UK Election 2017	1275	13%[c]	Direct Question	10%	[0.08; 0.12]	95%
		2554	13%[c]	List Method	14%	[0.06; 0.22]	87%
	Non-Voting New Zealand Election 2017	1717	6% [c]	Direct Question	4%	[0.03; 0.05]	98%
		3405	6% [c]	List Method	10%	[0.05; 0.15]	91%
Lamb and Stem (1978)	Whether Student Ever Failed a Class	63	30%	Direct Question	29%	[0.18; 0.4]	–[b]
		121	35%	RRT	36%	[0.19; 0.54]	–[b]
Locander et al. (1976)	Drunk Driving in Past Year	63	1	Direct Question (Telephone)	54%	[0.4; 0.68]	54%
		62		RRT	65%	[0.37; 0.93]	65%
	Bankruptcy in Past Year	60	1	Direct Question (Telephone)	71%	[0.56; 0.86]	71%
		55		RRT	100%	[1; 1]	100%
Markhof et al. (2025)	Tax Evasion by Firms in Uganda	1571	43%[c]	Direct Question	9%	[0.07; 0.11]	60%
		1320		BTS[d]	10%	[0.08; 0.12]	58%
		1459		List Method	20%	[0.16; 0.24]	59%
		1460		RRT	30%	[0.24; 0.36]	57%
Tracy and Fox (1981)	Number of Previous Arrests	120	1.78	Direct Question	1.02	[0.73; 1.27]	-[b]
		410	1.45	RRT	0.77	[0.44; 1.10]	-[b]
van der Heijden et al. (2000)	Social Security Fraud	99	1	Direct Question (Face to Face)	25%	[0.16; 0.34]	25%
		47		Direct Question (Computer)	19%	[0.08; 0.3]	19%
		96		RRT Variation 1	43%	[0.3; 0.56]	43%
		105		RRT Variation 2	49%	[0.33; 0.65]	49%
Wolter and Preisendörfer (2013)	Prior Conviction in Sample of Actual Convicts	219	1	Direct Question	58%	[0.51; 0.64]	58%
		332		RRT	60%	[0.51; 0.69]	60%

Notes: We extracted the percentages from the respective papers, rounding to a whole number. All papers have a sample in which they observe the true outcome at an individual level and can compare this to the individual's survey response. *95% CI* gives the 95% confidence interval for the estimate.

[a] In this experiment, the true benchmark is only available on the aggregate; however the setting allows for recovery of individual level validation metrics under very minor assumptions, which is why we include the experiment in our sample.

[b] Number not reported in the paper. [c] The paper only reports the aggregate administrative benchmark over the whole sample. [d] BTS = Bayesian Truth Serum

B Direct-Question Studies with Individual-Level Benchmarks

Appendix Table A2: Summary of Direct-Question Studies with Individual-Level Benchmarks

Reference	Outcome	Sample Size	True Value	Direct Question Estimate	Accuracy
Alem et al. (2018)[a]	Return Erroneously Received Money	156	32%	65%	48 %
Archambault et al. (2023)	Covid Vaccinated	1361	67%	69%	96%
Bekkers and Wiepking (2011)	Donation to Charity (Median)	105	15€	25€	-[b]
Chan et al. (2023)	Current Smoker (Female Soldiers Sample)	114	26%	19%	89%
Colley et al. (2018)	Minutes of Physical Activity	2372	23 mins	49 mins	-[b]
Dutz et al. (2021)	Applied for Unemployment Insurance (No Incentive Survey)	1700	9%[c]	8%	98%
Kleven (2022)	Non-Voting in Norway General Election 1969-2021	26333	14%	11%	96%
Kleven and Ringdal (2020)	Completion of Tertiary Education	5780	37%	43%	91%
Kvalvik et al. (2012)	Smoking During Pregnancy	2997	15%	13%	96%

Note: Percentages are rounded to the whole percentage point and decimals to a whole number.
[a] We use the following procedure to calculate prevalence and accuracy from the replication data: We pool over all experimental conditions. Following one of the main specifications of Alem et al. (2018), we exclude respondents stating they would "send some money back" in the survey, as this is difficult to compare with revealed behavior in their setup. As in the reported regressions, we use a dummy to assess whether the person actually sent money back, which is equal to one if at least some amount was sent back and zero if no money at all was returned.
[b] Accuracy and false positive rate is not strictly defined for non-binary outcomes.
[c] The paper only provides the true prevalence for the entire sample (not for the subsample of people in the no incentive survey); we thus report the true prevalence for the entire sample here.

C Additional Evidence for Prevalence of SDB (Raw Data)

Appendix Table A3: Summary of Accuracy, FPR and FNR by Technique for Studies Where All These Metrics Are Available.

Reference	Outcome	Accuracy	False Positive Rate	False Negative Rate
a. Direct Question				
Alem et al. (2018)[a]	Return Erroneously Received Money	48%	63%	30%
Archambault et al. (2023)	Covid Vaccinated	96%	8%	2%
Chan et al. (2023)	Current Smoker (Female Soldiers Sample)	89%	2%	33%
Höglinger and Jann (2018)	Cheating in Online Experiment 1 (Roll-a-six Game)	98%	1%	29%
Höglinger and Jann (2018)	Cheating in Online Experiment 2 (Prediction Game)	79%	0%	90%
Kleven (2022)	Non-Voting in Norway General Election 1969–2021	96%	26%	1%
Kleven and Ringdal (2020)	Completion of Tertiary Education	91%	12%	4%
Kuhn and Vivyan (2022)	Non-Voting UK Election 2017	95%	27%	2%
Kuhn and Vivyan (2022)	Non-Voting New Zealand Election 2017	98%	29%	0%
Kvalvik et al. (2012)	Smoking During Pregnancy	96%	1%	18%
Markhof et al. (2025)	Business Tax Evasion in Uganda	60%	5%	85%
b. List Method				
Kuhn and Vivyan (2022)	Non-Voting UK Election 2017	87%	43%	8%
Kuhn and Vivyan (2022)	Non-Voting New Zealand Election 2017	91%	39%	8%
Markhof et al. (2025)	Business Tax Evasion in Uganda	59%	16%	75%
c. RRT				
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 1 (Roll-a-six Game)	86%	12%	47%
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 2 (Prediction Game)	73%	11%	72%
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 1 (Roll-a-six Game)	95%	3%	45%
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 2 (Prediction Game)	78%	0%	85%
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 1 (Roll-a-six Game)	97%	-4%	59%
Höglinger and Jann (2018)[b]	Cheating in Online Experiment 2 (Prediction Game)	76%	-2%	91%
Markhof et al. (2025)	Business Tax Evasion in Uganda	57%	27%	65%

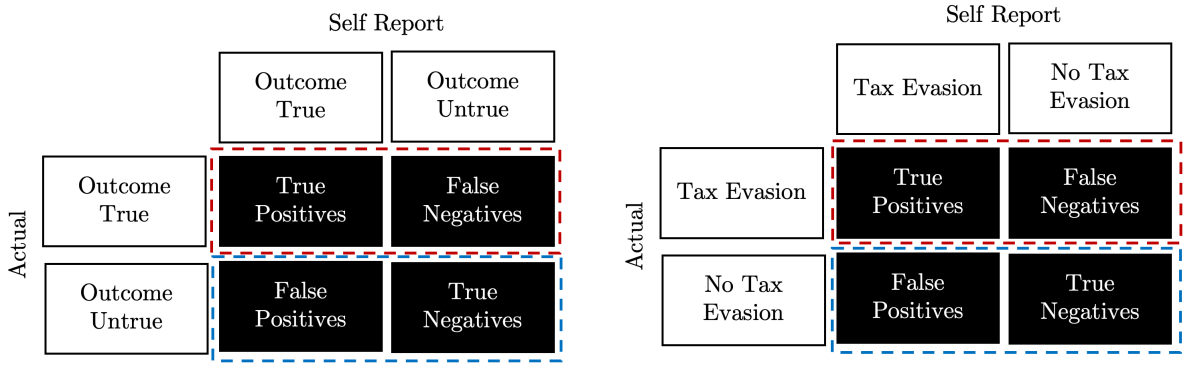
Note: For our metric definitions, see Appendix Section D.

[a] We use the following procedure to calculate prevalence and accuracy from the replication data: We pool over all experimental conditions. Following one of the main specifications of Alem et al. (2018), we exclude respondents stating they would "send some money back" in the survey, as this is difficult to compare with revealed behavior in their setup. As in the reported regressions, we use a dummy to assess whether the person actually sent money back, which is equal to one if at least some amount was sent back and zero if no money at all was returned.

[b] Höglinger and Jann (2018) test three different variations of the RRT for each of their games, we report estimates for every variation here.

D Definition of Metrics

We adopt the usual definitions for our metrics derived from the confusion matrix (Figure A1):



Box In Red: Data for Calculation of False Negative Rate

Box in Blue: Data for Calculation of False Positive Rate

Appendix Figure A1: General confusion matrix and illustrative example for tax evasion.

The overall *accuracy* measures the share of survey responses that correctly match the benchmark. Formally, it is the ratio of all correct classifications (true positives and true negatives) to the total number of cases in the confusion matrix,

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}}. \tag{1}$$

The false positive rate (FPR) is the ratio of people who wrongly self-report having performed a behavior divided by all people who actually have not performed the behavior,

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{FP} + \text{True Negatives (TN)}}. \tag{2}$$

For instance, if the behavior is tax evasion, the FPR is therefore the share of compliant taxpayers who report having evaded taxes among all compliant taxpayers. The false negative rate (FNR) is the ratio of people who wrongly self report *not* performing in a behavior divided by all people who actually performed the behavior,

$$\text{FNR} = \frac{\text{False Negatives (FN)}}{\text{FN} + \text{True Positives (TP)}}. \tag{3}$$

In the example tax evasion, the FPR is the share of compliant taxpayers who report having evaded taxes among all compliant taxpayers.