A Comprehensive Test of the Most Promising Method to Capture Social Desirability Bias in Online Surveys

Daniel Bischof *University of Münster & Aarhus University*
Tim Lars Allinger *Aarhus University*
Morgan Le Corre Juratic *Aarhus University*
Kristian Vrede Skaaning Frederiksen *Aarhus University*

August 1, 2024

Abstract

Social scientists have long debated the question of how much social desirability biases affect the information they can gather from online survey responses. However, it remains unclear how and to what extent we can measure it. Reviewing relevant literature focusing on this problem, we argue that the most promising way to measure social desirability bias is manipulating it globally through an experimental design placed at the very start of a survey. This approach—if successful—allows researchers to achieve three crucial goals that other approaches fall short in achieving simultaneously: 1) assuring that social desirability rather than confounders is measured, 2) allowing for checking whether social desirability was actually manipulated, and 3) allowing for measuring social desirability pressures in an infinite number of outcomes throughout the survey. Employing both novel treatment designs and designs already used in established research, we demonstrate with pre-registered survey experiments in the United States (N = 5,000) and Denmark (N = 3,000) that this approach is much too risky for researchers to pursue. Specifically, we show that some treatment designs repeatedly fail to achieve manipulation (i.e., respondents do not believe their answers are being observed), whereas others achieve manipulation but do not affect outcomes which we know for a fact are marred by social desirability (i.e., respondents do not care even if they know they are being observed). We end the paper by providing advice for scholars regarding which approaches are then most feasible to pursue judging by to what extent they achieve the three crucial goals reported above.

Department of Political Science, Aarhus University (DK); db@danbischof.com. We are thankful for comments by Francesco Colombo, Sirianne Dahlum, Florian Foos, Mashail Malik, Jonathan Mummolo, Brendan Nyhan, Erik Peterson, Vicente Valentim, Markus Wagner, and Natasha Wunsch. We thankfully acknowledge funding from the Carlsberg Foundation (Semper Ardens: Accelerate 2021) and the Aarhus University Research Foundation (AUFF Starting Grant 2021).

1 Introduction

Social scientists have always been concerned about whether or not they can accurately capture the concepts they are interested in. A major part of this concern comes from worries about social desirability biases (Maccoby and Maccoby 1954; Nederhof 1985; Tourangeau and Yan 2007)—namely, that social norms in societies lead citizens to answer survey items dishonestly. Some researchers have begun to view these biases not only as issues to be fixed but also as subjects worthy of study (see for instance: Jenkins et al. 2021; Valentim 2021; Colombo 2022; Bischof, Allinger, Le Corre Juratic, and Frederiksen 2023; Malik and Siddiqui 2024; Aycinena, Bogliacino, and Kimbrough 2024).

With the growing interest in detecting social desirability effects in surveys, the variety of designs used by scholars has also increased. However, substantial uncertainty remains as to whether and to what extent social desirability biases can actually be measured. In this paper, we review different approaches to measuring social desirability biases, define core goals for scholars attempting to do this, and provide a comprehensive test of the most promising approach based on the extent to which these goals are achieved.

We specifically argue that approaches to measure or remove social desirability biases should be judged based on the extent to which they achieve *three goals*. The first goal is assuring that social desirability rather than confounders are measured. Most experimental designs—including list experiments (Miller 1984)—aimed at measuring or removing social desirability achieve this goal. The second goal is allowing for checking whether and to what extent social desirability was actually manipulated. This requires a manipulation check following treatment—or more precisely the opportunity to include such a check. Most existing approaches can be combined with reliable manipulation tests, but list experiments—the most widely used by social scientists (Blair, Coppock, and Moor 2020)—fail this criterion. The third goal is maximizing the number of outcomes—or measures—that the approach allows for removing or measuring social desirability in. This is important as *a priori* researchers cannot perfectly know which items within their survey might be subject to social desirability biases.

To our knowledge, the only approach in existing social science research that has the potential to achieve all three goals is including a treatment aimed at manipulating an overarching sense of being observed by other people while being in a survey environment (see for instance: Bursztyn, Egorov, and Fiorin 2020; Colombo 2022). We therefore provide a comprehensive test of this "global treatment" approach, in which we employ four different treatments varying the observing reference group. Some of these treatments have been employed by prior research, while others are novel designs that mimic social desirability triggering characteristics identified by the literature. Specifically, we test all four treatments on a pre-registered sample of 5,000 participants in the United States and replicate the findings for the most promising one of them in a sample of 3,000 participants in Denmark—the reasoning being that we wanted to test the most promising treatment in the most likely context in terms of citizens being affected by social norms (see e.g. Torpe 2003).

First, we find that most treatments fail in actually manipulating perceptions of being observed. Second, none of the treatments successfully detect social desirability biases in constructs previously identified by research as prone to such biases *(e.g., donations, turnout, racial prejudice).* Hence, the most promising approach to measure social desirability bias in existing research seems too risky for researchers to pursue. Instead, we recommend that researchers prioritize among the stated goals to achieve at least two of them. This prioritization depends on the goals of individual research studies, but most often we believe that researchers will benefit from prioritizing the first two goals—causal identification as well as checking and assessing the strength of manipulation—at the expense of the third, the breadth of outcomes. The perhaps most prominent example of doing this comes from research on social norms in which the researcher assigns measure-specific vignettes manipulating social norms in regard to some matter to test whether manipulating information about other people's opinions affect survey responses in regard to that matter specifically (see for instance: Bicchieri et al. 2021; Bursztyn et al. 2023). This approach does not suffer from endogeneity problems and allows for including a manipulation check, but restricts the number of outcomes to one (or at least one per treatment).

This paper contributes to the knowledge on social desirability bias in survey research by providing a framework for assessing different approaches and—most importantly—by providing the most comprehensive test to date of the most promising "global treatment" approach (see for instance: Bursztyn, Egorov, and Fiorin 2020; Colombo 2022). Crucially, our findings do not imply that it *never* works, but rather that it is *too risky*—particularly in terms of economic costs—for researchers to pursue. Similarly, our findings do not imply that the approach would not work under certain circumstances; for example, with a very attentive sample. However, inattentiveness in surveys is a condition that robust approaches to measure social desirability need to be able to deal with (Ternovski et al. 2022). That said, robustness tests displaying heterogeneous treatment effects show that the global treatment approach does not prove effective for attentive respondents either (or any other relevant subgroup). Finally, our results align with previous findings suggesting a limited role of social desirability in online survey design and experiments (Kreuter, Presser, and Tourangeau 2008; Mummolo and Peterson 2019), but we do not argue that no such bias exists in online survey responses. Our findings show that attempting to reveal biases by inducing perceptions of being observed is likely to fail—either because respondents do not believe the treatment information or simply do not care about being observed in the context of an online survey.

2 Measuring Social Desirability: From the Bias to the Norm

The idea that social desirability affects what we can learn from survey data is hardly new. In fact, questions of if, and to what extent, social desirability forestalls the interpretation of survey answers as "truthful" beliefs and attitudes from respondents emerged with the first large-scale survey designs in the social sciences. More recently, researchers have shown a growing interest of social desirability in survey responses as a magnifying glass revealing social norms and their shifts in society. These two, albeit related, understandings of social desirability have led scholars to adopt different tools and measures to account for it in survey designs.

The first understanding of social desirability interpreted this phenomenon as a bias to be minimized in survey answers. In spite of revealing true preferences and attitudes in survey responses, researchers suspect that respondents may express the answer perceived to be "correct" and socially acceptable instead when being directly asked about sensitive topics (Maccoby and Maccoby 1954; Fisher 1993). The most prominent and early example of social desirability pressures was probably the "Hawthorne effect". Landsberger (1958) discovered within a study design that workers' productivity increased as long as they were observed and then decreased after their observation ended. Following this landmark example of the so-called "demand-effect" (Orne 1962; Mummolo and Peterson 2019), early studies in social sciences found regular instances of this social desirability bias, not only stemming from perceived researchers' expectations, but those of society at large. For instance, political scientists found that a highly socially desirable behavior, turnout, was and remains consistently over-reported (Campbell et al. 1960; Holbrook and Krosnick 2010). Beyond turnout, a large number of studies interested in sensitive behavior from a social perspective, such as racial and religious prejudice, corruption, or clientelism, uncovered similar under-reporting and mismatch between reported attitudes and actual behavior (Aronow et al. 2015; Corstange 2018; Gonzalez-Ocantos et al. 2012; De Jonge 2015).

Given this approach, most of this research has constantly sought to develop designs minimizing these social desirability pressures to accurately measure socially sensitive attitudes and behavior. The most well-known fixes for social desirability bias are arguably the list experiment (also referred to as the "item count technique"; see Miller 1984), which consists of estimating the severity of the bias via the random assignment of a list of statements including or excluding the sensitive item. Other approaches common in the literature are the randomized response technique (Warner 1965) and the cross-wise technique (Yu, Tian, and Tang 2008; for a review, see Gingerich et al. 2016). Lately, the increased relevance of conjoint survey designs has been understood to be largely unaffected by social desirability bias (Hainmueller, Hopkins, and Yamamoto 2014; Horiuchi, Markovich, and Yamamoto 2022). However, these approaches share a common limitation. These studies aim to reduce social desirability bias by design, but have the limitation that they do not allow for including manipulation checks to learn whether social desirability pressure is actually being induced. Recent studies have found that inattention, education level, or more generally simple errors and non-strategic misreporting may explain these differences instead of a social desirability effect (Castro Cornejo and Beltrán 2012; Kramon and Weghorst 2019; Kuhn and Vivyan 2022). We therefore label these approaches *"experimental black box"* approaches.

A second perspective adopted more recently by social scientists argues that because it reveals social pressures derived from established social norms in society, social desirability deserves more attention than merely being trimmed away by survey design. These approaches consider that social desirability effects will typically reflect an adaptation to perceived appropriate attitudes or behavior, and scholars have therefore attempted to manipulate these social pressures to reveal norms in society. Social scientists have, among others, used this perspective on social desirability to study social norms on fairness (Bicchieri and Chavez 2010), gender (Bursztyn et al. 2023), or on the role of norms in shifting stigmatized, authoritarian, and xenophobic attitudes (Bursztyn, Egorov, and Fiorin 2020; Colombo 2022; Valentim 2024).

The norm-based approaches typically come in two overarching versions in survey designs. The first version manipulates social norms directly in a measure-specific vignette to check whether certain behaviors or attitudes are affected by information about societal behaviour (or descriptive and normative expectations, see Bicchieri (2016)). Social desirability, in this case, is measured via the conditionality of respondents' attitudes and behavior to this information about societal norms. A typical example comes from Bicchieri et al. (2021) who assign vignettes informing respondents about support for social distancing among other citizens to test whether citizens' own attitudes are affected by social norms, hence being vulnerable to social desirability pressures. This approach is rather promising, as it allows for both random assignment of norms as well as checking whether manipulation was successful: *did perceived norms actually shift?* The main drawback is that only one measure/outcome can be measured per treatment by design.

The other approach employs a more "global" strategy, where social desirability pressures are manipulated in a way by randomly varying the extent to which respondents' answers are observed or remain private. For example, Bursztyn, Egorov, and Fiorin (2020) manipulate the wording of the confidentiality statement to suggest that respondents' answers will be made public on a local website to assess shifts in xenophobic norms following the election of Donald Trump through a vignette experiment. Colombo (2022) adopts a similar manipulation of confidentiality in a vignette prior to his outcome of interest. This second, global version provides the most promising avenue for researchers to measure the extent of social desirability in survey data by allowing to accomplish all three important objectives. By manipulating the perceived publicity of individual survey responses and not only ways to remove it as bias, this approach allows for causal identification (Goal 1) as well as allowing for assessment of the strength of manipulation (Goal 2). In addition, a yet under-explored avenue is the use of such treatment to assess social desirability pressures in an efficient way on a wide range of outcomes of interest to researchers (Goal 3) as opposed to being constrained on a single sensitive outcome per treatment by design.

Table 1 sums up the approaches. One approach that we—for good reason—have not touched upon yet is using "direct" measures of social desirability (i.e., measures not using random assignment see e.g. Crowne and Marlowe 1960; Martin 1984). These approaches include social desirability scales, single measures, etc., where researchers check whether there are heterogeneity in responses across how prone to social desirability respondents are (e.g., are respondents scoring high on social desirability scales more likely to report that they turned out at elections?). "Black box" approaches then include list experiments, cross-wise techniques, conjoint experiments etc., which aim to reduce social desirability bias using random assignment on a single outcome but do not allow to check whether and to what extent social desirability was manipulated. Vignette treatments—which also could be termed "specific scenario vignette" treatments—allow for causal identification as well as checking manipulation, but are limited to one measure per treatment. Finally, global approaches achieve all three goals. Importantly, this does not mean that they necessarily *do* manipulate social desirability or captures such pressures on many outcomes—the global approach just allows for actually checking for these things in the first place.

Table 1: Overview of approaches

approach	causal (1)	check (2)	outcomes (3)
direct measure	no	no	yes
black box treatment	yes	no	no
vignette treatment	yes	yes	no
global treatment	yes	yes	yes

However, a systematic study assessing and comparing different types of treatments instigating the perception of being observed by respondents is still lacking in the current literature. Yet, this assessment remains necessary, as the current treatment designs manipulating respondents' perceptions of the privacy of their answers vary in important ways. Most notably, social desirability pressures may be manipulated using different observing reference groups, such as ordinary citizens, researchers, or the wider public (Bursztyn, Egorov, and Fiorin 2020; Bicchieri et al. 2021; Colombo 2022). In addition, some compound effects of being observed, such as anxiety or fear of social sanctions, may increase dropout rates, thereby reducing the validity of the approach. This study therefore aims to contribute to the current literature by providing a systematic analysis of the global treatment approach by including different treatment versions in the same survey. These treatments include both those used in existing research and novel ones developed by us. Specifically, we design an online survey experiment manipulating the extent to which respondents feel that their answers are being observed with variation on the observer group and assess efficiency in terms of manipulation, compound effects, dropout rates, and validity using survey questions known to be marred by social desirability bias.

3 Experimental design

To achieve this goal and systematically evaluate the quality of social desirability treatments, we designed a survey experiment containing four social desirability treatments and a control group. We teamed up with Cint to get access to a representative sample of American citizens. We conducted the first stage of our fieldwork on a sample of 3,000 Americans in June-July 2023, testing and fine-tuning the design of our pre-analysis plan (Bischof, Allinger, Juratic, Frederiksen, and Valentim 2023) [1]. Our main fieldwork took place on a new representative sample of 5000 American respondents, followed by another test on another 3,000 Danish citizens to enhance the external validity of our findings and test further the most promising of these treatments.

3.1 Treatment design

We designed an experiment to assess a) how well different treatments on social desirability work *(manipulation)* and b) what they allow us to reveal in terms of such biases on a range of outcomes *(treatment effects)*. Table 2 gives an overview of the treatments we are using in our experiment.

Table 2: Overview of treatments used in our study

"treatment"	reference group	attention needed	potential dropouts	likely compound effects
confidentiality	public	high	high	anxiety
anonymity	public	medium	low	confidence, trust
chatbot	experimenter	low	medium	anxiety
pairing	fellow citizens	medium	medium	anxiety

All four treatments aim at inducing (or reducing) feelings that respondents' answers will be observed. Our first treatment, *confidentiality*, stresses that: “The results from this survey, including your individual opinions, will be posted on our website in approximately one month after the analysis is completed. We will notify you when the results become available on our website (website link)” (For

[1] All deviations from PAP can be found in SI.6.1

a similar approach, see: Bursztyn, Egorov, and Fiorin 2020; Colombo 2022).[2] The second treatment, *anonymity*, seeks to achieve the exact opposite of the confidentiality treatment by re-assuring respondents with a brief statement on a single page following the consent form that "at no point will their individual responses" be publicly accessible. The third treatment, *chatbot*, is a pop-up revealing the following information to respondents: "Hi! I am the research assistant of this project and I will be live with you while you respond to our survey. If any question emerges, please do not hesitate to get in touch by writing to me at the following address: [contact address], and I will come back to you shortly." This treatment aims to induce social desirability pressures via the interviewer or experimenter demand effects (EDE), in line with potential studies assessing social desirability biases (Ejaz and Thornton 2023; Mummolo and Peterson 2019; Orne 1962; Jenkins et al. 2021; Valentim and Widmann 2023). The fourth treatment, *pairing*, mimic approaches using observational evidence (Valentim 2024) or behavioral games (Bicchieri and Chavez 2010) where the reference group or observer becomes another ordinary respondent. This treatment, following consent, outlines that "at the end of the survey responses might be shown to other respondents just like you".

It is important to note that these treatments, while all putting emphasis that answers may be viewed by other people, vary on at least four different important aspects beyond the suggested observer (also reported in the columns of Table 2).

First, in our judgment they vary in terms of the amount of attention needed by respondents to pick up the treatment. Confidentiality treatments, as also designed by other researchers in the field (Bursztyn, Egorov, and Fiorin 2020; Colombo 2022), appear within a text with general information on respondents' data treatment and anonymization, information that most respondents part of online samples are used to and rarely reveal any new information to these "professionals". This means that many respondents will not spend much time on this screen but instead scan the page quickly before proceeding. Put differently, many respondents will not pay enough attention to pick up the treatment. In contrast, the remaining treatments are very loud, sometimes even interrupting (chatbot), and it seems unlikely that respondents miss the information *a priori*. Besides attention to the treatments also vary in terms of whether observation must be inferred by respondents (i.e. chatbot), or whether

[2] One might stress that putting this treatment on a separate screen might increase the manipulation effects. We refrain from such a design to ensure that our findings replicate previous approaches as closely as possible. Yet, one can understand the anonymity treatment as doing exactly that; respondents receive this information at a separate screen and effects only partly differ as reported later.

it is definitive (i.e., the remaining treatments) as well as in how “explicit” they are.

Second, these treatments, if successful, might raise respondents’ suspicion. They might fear observation by others and, thus, simply drop out in much higher rates than in usual survey environments. This risk is most likely the highest for people who pick up the confidentiality treatment and lowest for the anonymity treatment.[3]

Moreover, the treatments might create various, and most importantly, divergent compound effects which we outline in the last column of Table 2. This is highly relevant as these treatments seek to induce social desirability pressures. However, such pressures might come along with feelings of anxiety or conversely, increased trust and confidence. It is clear that the four treatments tend to reduce compound treatments in comparison to observational studies seeking to study social desirability via interview modes, but differences across the treatments might still exist and result in divergent treatment effects.

Finally, it is worthwhile mentioning that all treatments except anonymity use mild deception. We assess deception as mild, because it only occurs regarding perceptions of what happens in relation to the survey—the treatments do not manipulate or misguide perceptions of general societal matters. We also handle this using debriefing informing respondents that and how deception was used at the end of the survey.

3.2 Outcomes & manipulation tests

A key advantage of global social desirability treatments is that the potential effects of social desirability can be tested on a range of outcomes. In contrast, other approaches usually need to focus on one outcome which is directly linked to a treatment—such as fictitious vignette scenarios (e.g. Bicchieri et al. 2021) or the ‘bogus pipeline’ approach (e.g. Hammer, Banks, and White 2014) .

Nevertheless, we selected a range of outcomes based on the idea that either previous research reported strong effects of social desirability, combining either over or undereporting behavior (i.e. turnout, donations, racism, see for instance classical work by Gerber, Green, and Larimer (2008); Karp and Brockington (2005); Blair, Coppock, and Moor (2020)), or for which we have historical reasons to assume that some socially desirable behavior seems obvious (i.e. support for communism in the US

[3] However, respondents might also become more suspicious if told repeatedly that their response will the treated with high confidence: why emphasize this if this is the common standard in modern survey research?

context). Of course it is difficult to know if social desirability biases do exist on these outcomes; as non-significant result might not have been brought forward to the same extent as significant findings *(file drawer problems)*. Nevertheless, it is reasonable to expect some social desirability biases for some of these outcomes. In total we created fourteen items for which we test whether differences due to treatments of social desirability become apparent.

Besides these outcomes we also asked respondents about their feelings (anger, disgust, enthusiasm, happiness, fear, and anxiety) while responding to our survey as well as several items that previous research have used to *directly* measure social desirability traits (Martin 1984). In addition, we test for possible problems with compoundedness by including a "provocation" indicator—specifically a need for chaos measure (Petersen, Osmundsen, and Arceneaux 2023)—and an indicator of trust in scientists to assess whether our social desirability treatment creates a form of backlash answer from respondents who feel observed, particularly so when primed with the observation by researchers. Finally, we also include a social desirability scale statement to rule out that our treatment does not affect respondents' psychological trait of being prone to social desirability pressures. This all highlights the degree to which our approach enables us to assess which treatment best induces social desirability pressures—and to what extent the treatment manipulates different aspects of social pressures.

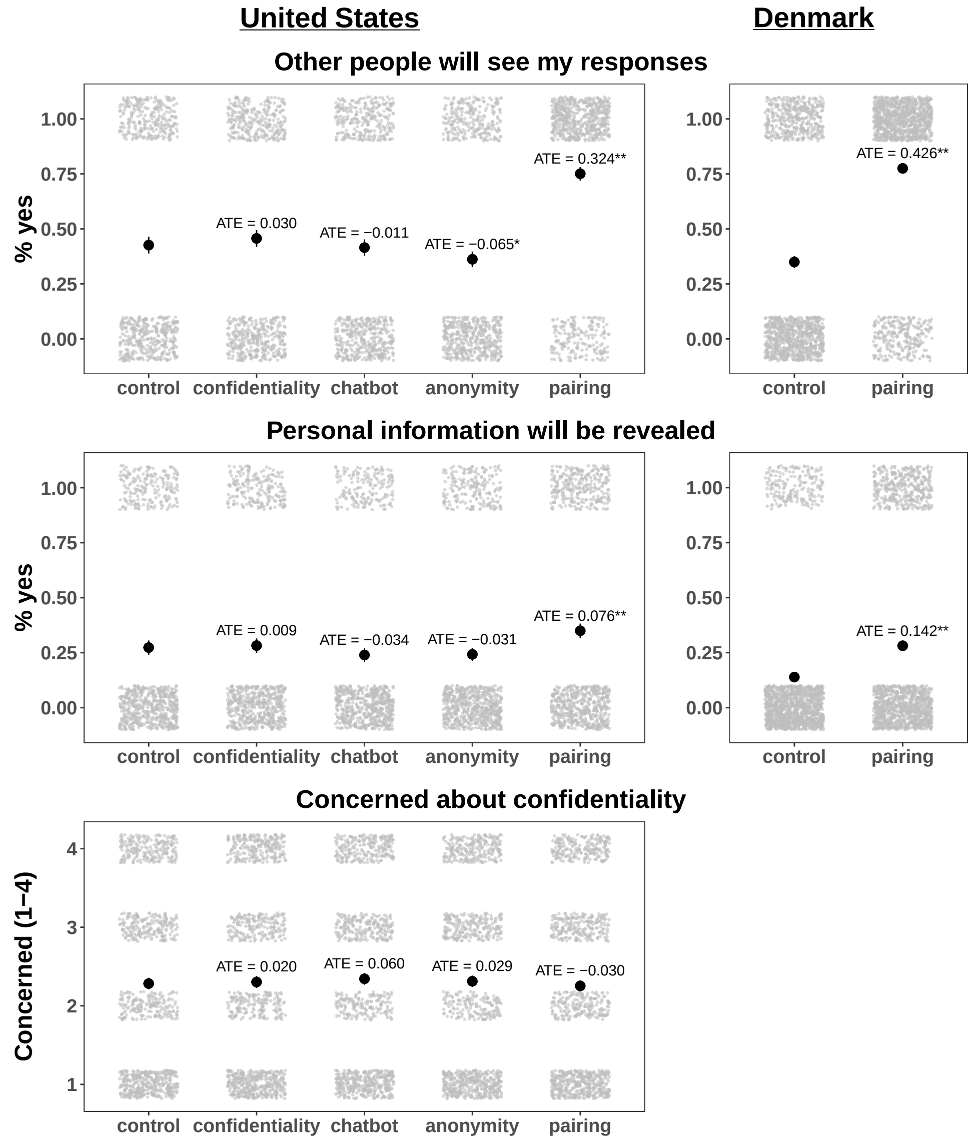
Another core advantage of the global approach is the option to include manipulation checks to test whether social desirability pressures were actually manipulated. We use the fieldwork of the American case to test how well each of these treatments work by employing four manipulation checks. The first manipulation check measures whether the respondents think that *other people* beyond the researchers will see their responses (as a matter of fact). The second check seeks to understand how *concerned* the respondents are that other people will see their responses. Using the same formulation as Bursztyn, Egorov, and Fiorin (2020), the third check quantifies how concerned the respondents are about the *confidentiality* of their responses (in the context of delivering data to researchers such as us). We included these different measures to avoid favoring specific treatments (i.e., including only a single measure which is more favorable to one treatment rather than others would skew the test).

4 Results

4.1 Manipulation

The key question for the following analysis is whether or not our treatments induced feelings of social desirability. Recall that what we seek to manipulate is how strong these feelings are for respondents in the treatment conditions in comparison to the control group. Thus, the research design idea is to induce a stronger (or weaker) feeling of social desirability pressure; the goal is not to have a control group “free” of any such biases—which is plausibly impossible to achieve.

Figure 1: Treatment effects on key manipulation checks



Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).
$* p \leq 0.05$, $** p \leq 0.01$

Figure 1 summarizes our findings for our key manipulation tests.[4] All in all, the findings are quite striking: we mostly report null findings. The confidentiality treatment reveals no manipulation effects; even not for the outcome directly designed to pick up such effects ("being concerned about confidentiality"). The only two treatments revealing some significant findings are the anonymity treatment—which seeks to reduce feelings of being observed and does so as indicated by its negative effect on "others will see my responses"—as well as the pairing treatment. For pairing we find particularly strong manipulation effects: a 32.4 percentage point increase in thinking that others will see respondents' answers. As mentioned above, we attempted to replicate the workings of this treatment in Denmark, where manipulation is even stronger (42.6 percentage points). Respondents also become much more likely to believe that some of their personal information will be revealed (7.6 percentage points increase in the US, 14.2 percentage points increase in Denmark). All in all the pairing treatment quite clearly outperforms all other treatments which barely have any sort of manipulation effects in our experiment.

4.2 Drop-outs, attention and other potential compound treatment effects

As discussed above, a key concern with increasing feeling of social desirability might be drop-outs: respondents who are concerned about their confidentiality might simply leave the survey upon receiving one of our treatments. However, as we show in the SI our treatment conditions do not report any significant differences in terms of dropouts from the survey (see Figure SI 3).[5]

As an additional concern, readers might suggest that the treatments, most specifically the chatbot treatment as a pop-up, might increase respondents' attention as a compound treatment effect. Yet again, we do not find any significant differences in a post-treatment attention test across our treatment conditions.

Similarly, it seems likely that feelings of anxiety and/or anger might be activated by e.g. the confidentiality treatment. In the SI we also report effects on how respondents feel while answering to our questions. However, we do not find any effects on emotions of any of our treatments (see

[4] We asked the key manipulation checks early in our survey design to ensure that any effects could be captured. Only covariates (gender, age etc.) were asked between treatment and manipulation checks (for a full overview of the survey flow please consult: Figure SI 1).

[5] As respondents receive all our treatments after their consent (except for the confidentiality), there can not be a plausible effect of the treatments on "no participation".

Figure SI 2).

Finally, we also do not find significant differences in how provoked respondents are by our treatments or how much they trust scientists. Both might be compound treatment effects that render our design incapable of providing evidence for socially desirable behavior (see Figure SI 3 in the SI).

4.3 Findings: null effects

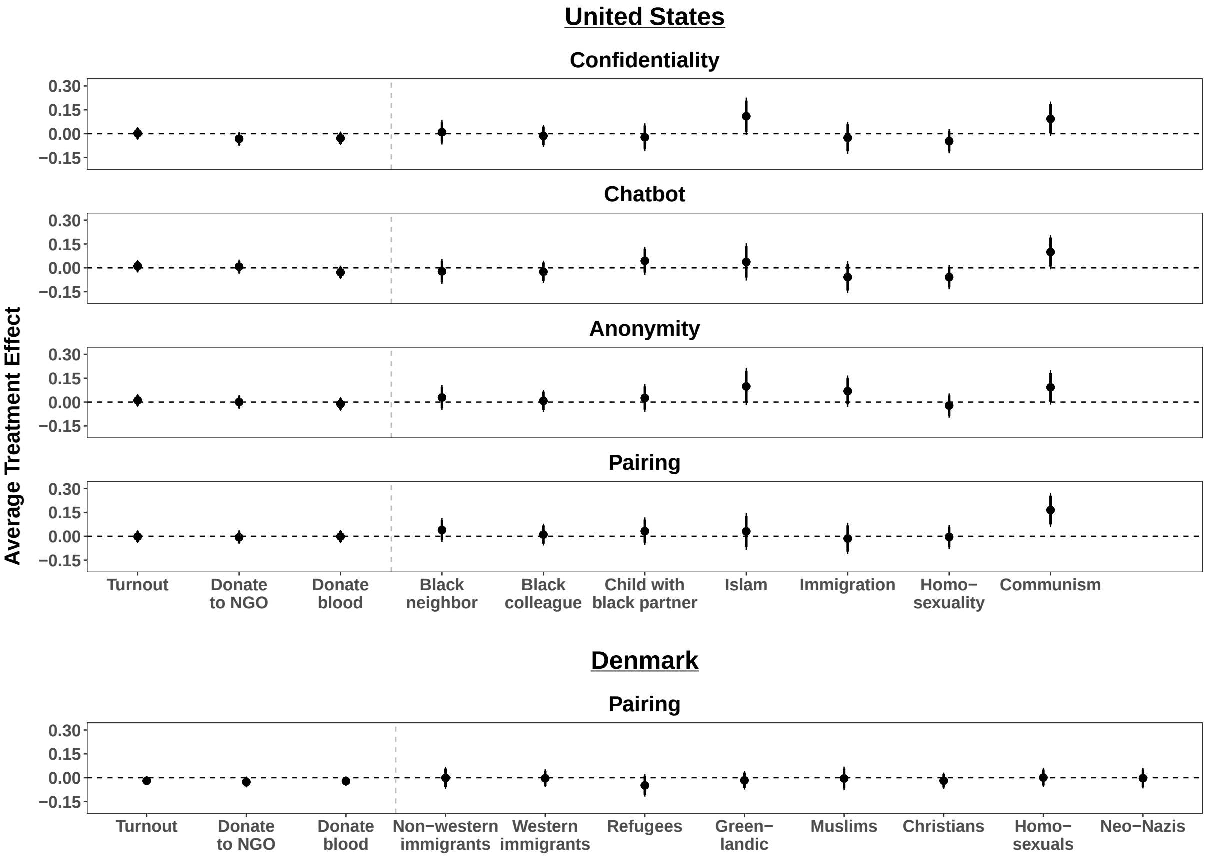
But how do the treatments affect likely outcomes subject to social desirability bias? Figure 2 reports our findings for both cases across ten/eleven different outcomes. We do not find *any* significant effects from our treatments across the outcomes—the only exception being support for communism in the US. However, given the amount of tests we are running this might as well we be subject to pure chance. Specifically, as all treatments seem to have effects on this outcome (several of which did not manipulate social desirability pressure), our interpretation is that the control group by random chance is more supportive of communism compared to the remaining groups.[6]

Notice that these null findings are not subject to any power constraints. Confidence intervals across the board are small and we pre-registered our study to capture very small effects of 4 percentage points (0.04) with 0.8 power. There is simply no effect of our treatments on such outcomes as turnout, donations, or racism.

One might suggest that many of these effects should not play out on the entire sample but instead for subgroups. For instance, it seems likely that racism mainly is subject to socially desirable behavior among whites. Similarly, communism might be an entirely different affair for Republicans than Democrats in the US. However, our heterogeneous treatment tests in the SI do not show such differences; there is again no social desirability detected for sub-populations (Figure SI 8 and Figure SI 9).

[6] Especially the fact that the direction and magnitude of the treatment effect in the Anonymity treatment equal those of all other treatments, even though the manipulation of the Anonymity treatment goes in the opposite direction, suggests that this finding is most likely to be noise.

Figure 2: Treatment effects on outcomes



Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

5 Conclusion and discussion

Approaches aiming to measure or to remove social desirability biases should be judged on basis of whether they achieve three goals: causally identifying social desirability, credibly checking whether social desirability was manipulated, and maximizing the number of outcomes in which social desirability can be measured. The most promising approach to achieve these goals is inducing feelings of social desirability in the start of a survey—the "global treatment" approach.

Leveraging two survey experiments in the United States and Denmark, we have provided the most comprehensive test to date of such an approach and demonstrated that it is much too risky for researchers to pursue. Global treatments often either fail in inducing feels of social desirability altogether or capturing social desirability in outcomes which we know are marred by it. Our choice of cases testify to this conclusion, as the global treatment approach fails not only in the United States but also in the strong social norms context of Denmark (see e.g. Torpe 2003). Importantly, the approach might sometimes be successful, but we show that it often is not, which especially means that risks in terms of economic costs are high.

One interpretation of why global treatments often fail is that there are strong limits to which reference groups the researcher can credibly build in to such treatments in online surveys. Based on our findings, it seems that the reference group—among our treatments—inducing most social desirability is ordinary citizens. Assigning more specific groups like family or neighbors might very well induce even stronger feelings of social desirability and yield a stronger treatment, but doing so is not possible to do credibly in a survey experiment (for doing so credibly in a field experiment, see Gerber, Green, and Larimer (2008)). It seems very unlikely that citizens would trust claims that researchers will manage to reveal their responses to such groups; for ethical as well as practical reasons. However, revealing information to other ordinary citizens is credible, as researchers do have the option to reveal responses to other individuals within the sample.

What should researchers wishing to capture social desirability biases then do? Our recommendation is re-considering the three goals mentioned above and prioritizing among them. In many cases, we suspect that researchers will benefit from prioritizing the first two goals (causal inference and checking manipulation) at the expense of the third (many outcomes). One way to implement this is using vignettes (see “vignette treatment” in Table 1) assigning information about what other people think about a behavior or topic of interest and subsequently measure how respondents’ attitudes or behaviors are affected by this information. A good example is Bicchieri et al. (2021) who measure to what extent social distancing attitudes are dependent on other people’s attitudes. This approach allows for random treatment assignment as well as checking whether respondents actually took in the information assigned, thus achieving goal one and two. This is a sobering recommendation, but it will at least ensure that researchers have a strong sense of what is going on in their survey and provide the possibility to document that.

References

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. “Combining list experiment and direct question estimates of sensitive behavior prevalence.” *Journal of Survey Statistics and Methodology* 3 (1): 43–66.

Aycinena, Diego, Francesco Bogliacino, and Erik O. Kimbrough. 2024. “Measuring norms: Assessing the threat of social desirability bias to the Bicchieri and Xiao elicitation method." *Journal of Economic Behavior & Organization* 222: 225–239.

Bicchieri, Cristina. 2016. *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford: Oxford University Press.

Bicchieri, Cristina, and Alex Chavez. 2010. "Behaving as expected: Public information and fairness norms." *Journal of Behavioral Decision Making* 23 (2): 161–178.

Bicchieri, Cristina, Enrique Fatas, Abraham Aldama, AndrÃ©s Casas, Ishwari Deshpande, Mariagiulia Lauro, Cristina Parilli, Max Spohn, Paula Pereira, and Ruiling Wen. 2021. "In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic." *PLOS ONE* 16 (6): e0252892.

Bischof, Daniel, Tim L Allinger, Morgan Le Corre Juratic, and Kristian V S Frederiksen. 2023. "(Mis-)Perceiving Support for Democracy: The Role of Social Norms for Democracies.".
URL: *osf.io/dpq7w*

Bischof, Daniel, Tim Lars Allinger, Morgan Le Corre Juratic, Kristian Vrede Skaaning Frederiksen, and Vicente Valentim. 2023. "Social Desirability Bias as Substance and Not Nuisance.".

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

Bursztyn, Leonardo, Alexander W Cappelen, Bertil Tungodden, Alessandra Voena, and David H Yanagizawa-Drott. 2023. "How Are Gender Norms Perceived?".

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From extreme to mainstream: The erosion of social norms." *American Economic Review* 110 (11): 3522–3548.

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E. Stokes. 1960. *The American Voter.* Chicago: University of Chicago Press.

Castro Cornejo, Rodrigo, and Ulises Beltrán. 2012. List Experiments, Political Sophistication, and Vote Buying: Experimental Evidence from Mexico. Technical Report 2.

Colombo, Francesco. 2022. "Collective Memory and the Stigmatization of Authoritarian Nostalgia: Evidence from Italy." *SSRN Electronic Journal* .

Corstange, Daniel. 2018. "Clientelism in Competitive and Uncompetitive Elections." *Comparative Political Studies* 51 (1): 76–104.

References

Crowne, Douglas P., and David Marlowe. 1960. "A new scale of social desirability independent of psychopathology." *Journal of Consulting Psychology* 24 (4): 349–354.

De Jonge, Chad P Kiewiet. 2015. "Who Lies About Electoral Gifts? Experimental Evidence from Latin America." *The Public Opinion Quarterly* 79 (3): 710–739.

Ejaz, Hamad, and Judd R. Thornton. 2023. "Survey mode and satisfaction with democracy." *Political Science Research and Methods* pp. 1–8.

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

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review* 102 (1): 33–48.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho, and Mauricio Ruiz-Vega. 2016. "When to protect? Using the crosswise model to integrate protected and direct responses in surveys of sensitive behavior." *Political Analysis* 24 (2): 132–156.

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Meléndez, Javier Osorio, and David W. Nickerson. 2012. "Vote buying and social desirability bias: Experimental evidence from Nicaragua." *American Journal of Political Science* 56 (1): 202–217.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22 (1): 1–30.

Hammer, Michael J, Antoine J Banks, and Ismail K White. 2014. "Experiments to Reduce the Over-Reporting of Voting: A Pipeline to the Truth." *Political Analysis* 22 (1): 130–141.

Holbrook, Allyson L, and Jon A Krosnick. 2010. "Social desirability bias in voter turnout reports: Tests using the item count technique." *Public Opinion Quarterly* 74 (1): 37–67.

Horiuchi, Yusaku, Zachary Markovich, and Teppei Yamamoto. 2022. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 30 (4): 535–549.

Jenkins, Clinton, Ismail White, Michael Hanmer, and Antoine Banks. 2021. "Vote Overreporting While Black: Identifying the Mechanism Behind Black Survey Respondents' Vote Overreporting." *American Politics Research* 49 (5): 439–451.

Karp, Jeffrey A, and David Brockington. 2005. "Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries." *The Journal of Politics* 67 (3): 825–840.

Kramon, Eric, and Keith Weghorst. 2019. "(Mis)Measuring Sensitive Attitudes with the List Experiment." *Public Opinion Quarterly* 83: 236–263.

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

Kuhn, Patrick M., and Nick Vivyan. 2022. "The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." *Political Analysis* 30 (3): 381–402.

Landsberger, Henry A. 1958. "Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry.".

Maccoby, Eleanor E, and Nathan Maccoby. 1954. "The Interview: A Tool of Social Science." In *Handbook of Social Psychology 1.* pp. 449–487.

Malik, Mashail, and Niloufer Siddiqui. 2024. "Third Party Presence and the Political Salience of Ethnicity in Survey Data." *The Journal of Politics* 86 (1): 364–368.

Martin, Harry J. 1984. "A Revised Measure of Approval Motivation and Its Relationship to Social Desirability." *Journal of Personality Assessment* 48 (5): 508–519.

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

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113 (2): 517–529.

Nederhof, Anton J. 1985. "Methods of Coping with Social Desirability Bias: A Review." *European Journal of Social Psychology* 15 (3): 263–280.

Orne, Martin T. 1962. "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American Psychologist* 17 (11): 776–783.

Petersen, Michael Bang, Mathias Osmundsen, and Kevin Arceneaux. 2023. "The "Need for Chaos" and Motivations to Share Hostile Political Rumors." *American Political Science Review* 117 (4): 1486–1505.

Ternovski, John, Lilla Orr, Joshua Kalla, and Peter Aronow. 2022. "A Note on Increases in Inattentive Online Survey-Takers Since 2020." *Journal of Quantitative Description: Digital Media* 2: 1–35.

Torpe, Lars. 2003. "Social capital in Denmark: A deviant case?" *Scandinavian Political Studies* 26 (1): 27–48.

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

Valentim, Vicente. 2021. "Parliamentary Representation and the Normalization of Radical Right Support." *Comparative Political Studies* 54 (14): 2475–2511.

Valentim, Vicente. 2024. "Political Stigma and Preference Falsification: Theory and Observational Evidence." *The Journal of Politics* .

Valentim, Vicente, and Tobias Widmann. 2023. "Does Radical-Right Success Make the Political Debate More Negative? Evidence from Emotional Rhetoric in German." *Political Behavior* 45 (1): 243–264.

Warner, Stanley L. 1965. "Randomized response: A survey technique for eliminating evasive answer bias." *American Statistical Association* 60 (309): 63–69.

Yu, Jun Wu, Guo Liang Tian, and Man Lai Tang. 2008. "Two new models for survey sampling with sensitive characteristic: Design and analysis." *Metrika* 67 (3): 251–263.

Supporting Information:

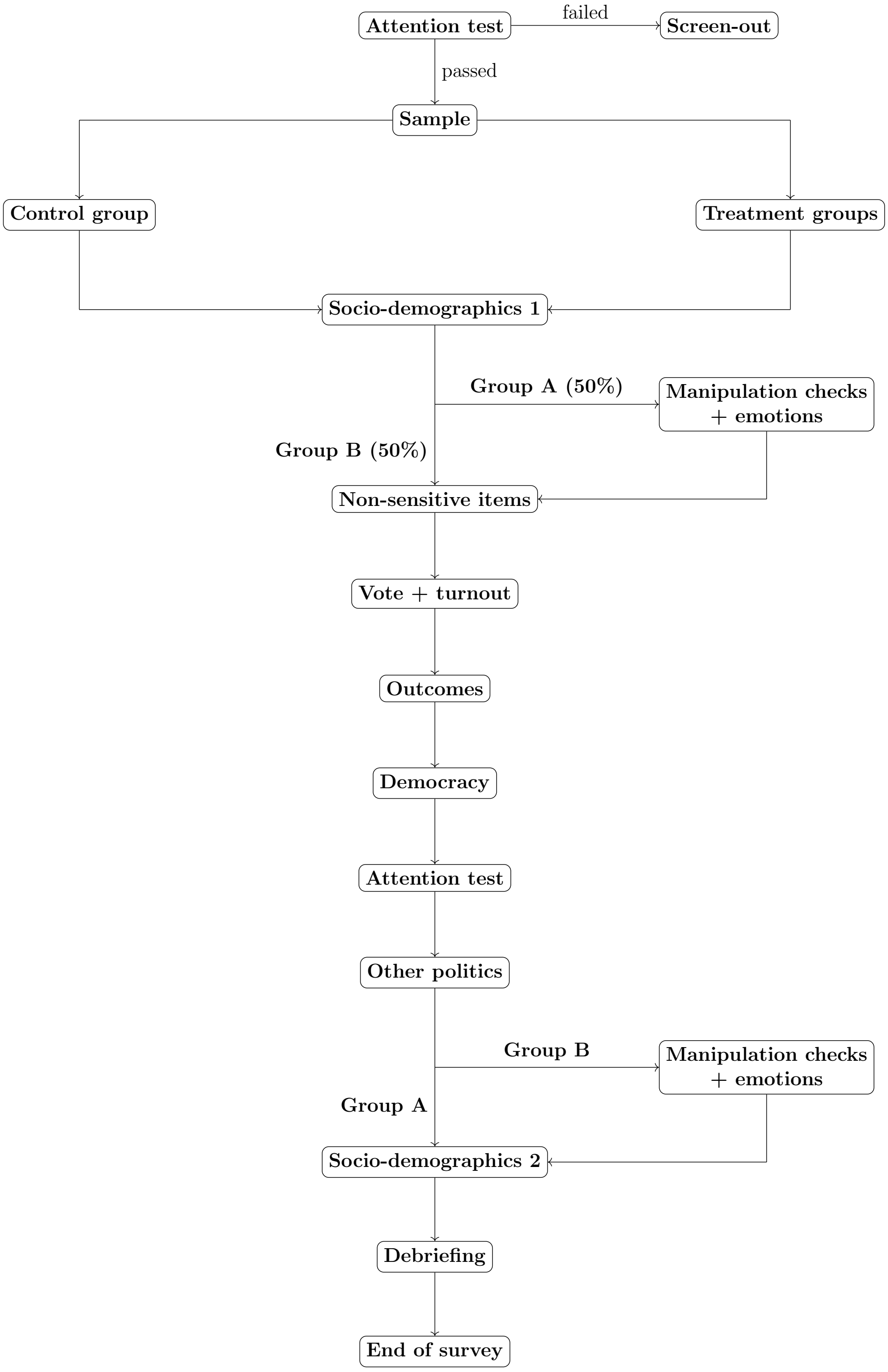
A Comprehensive Test of the Most Promising Method to Capture Social Desirability Bias in Online Surveys

SI Supporting Information

Contents

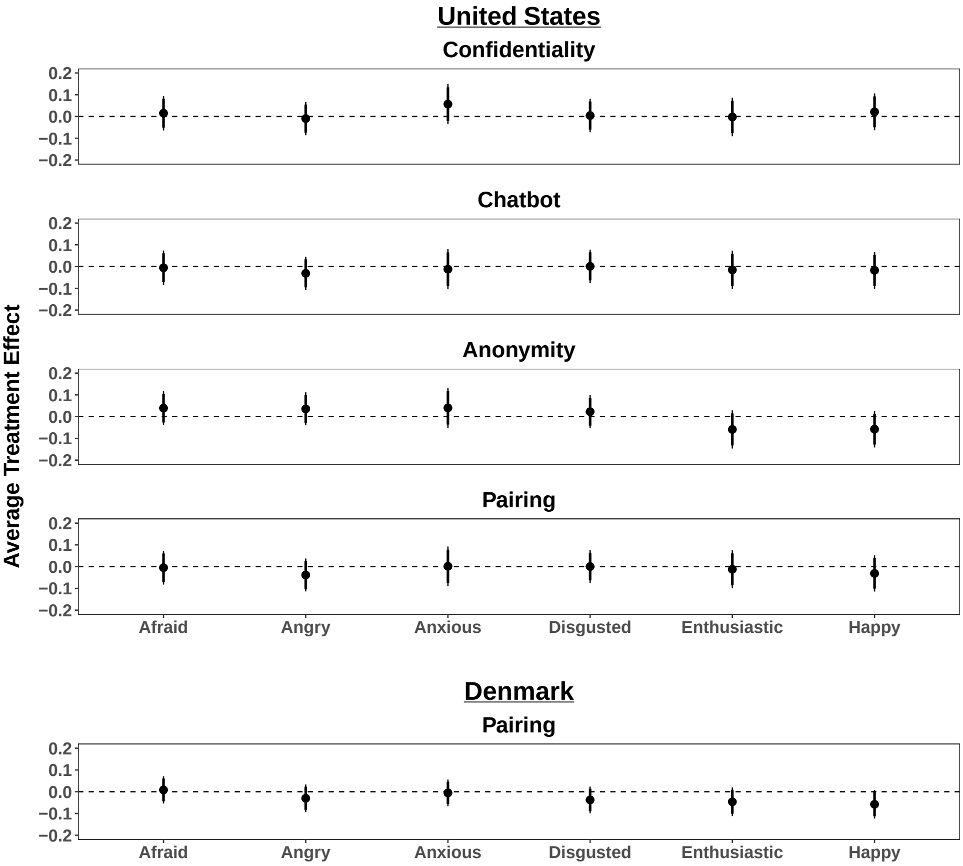
SI.1 Survey flow

Figure SI 1: Survey flow



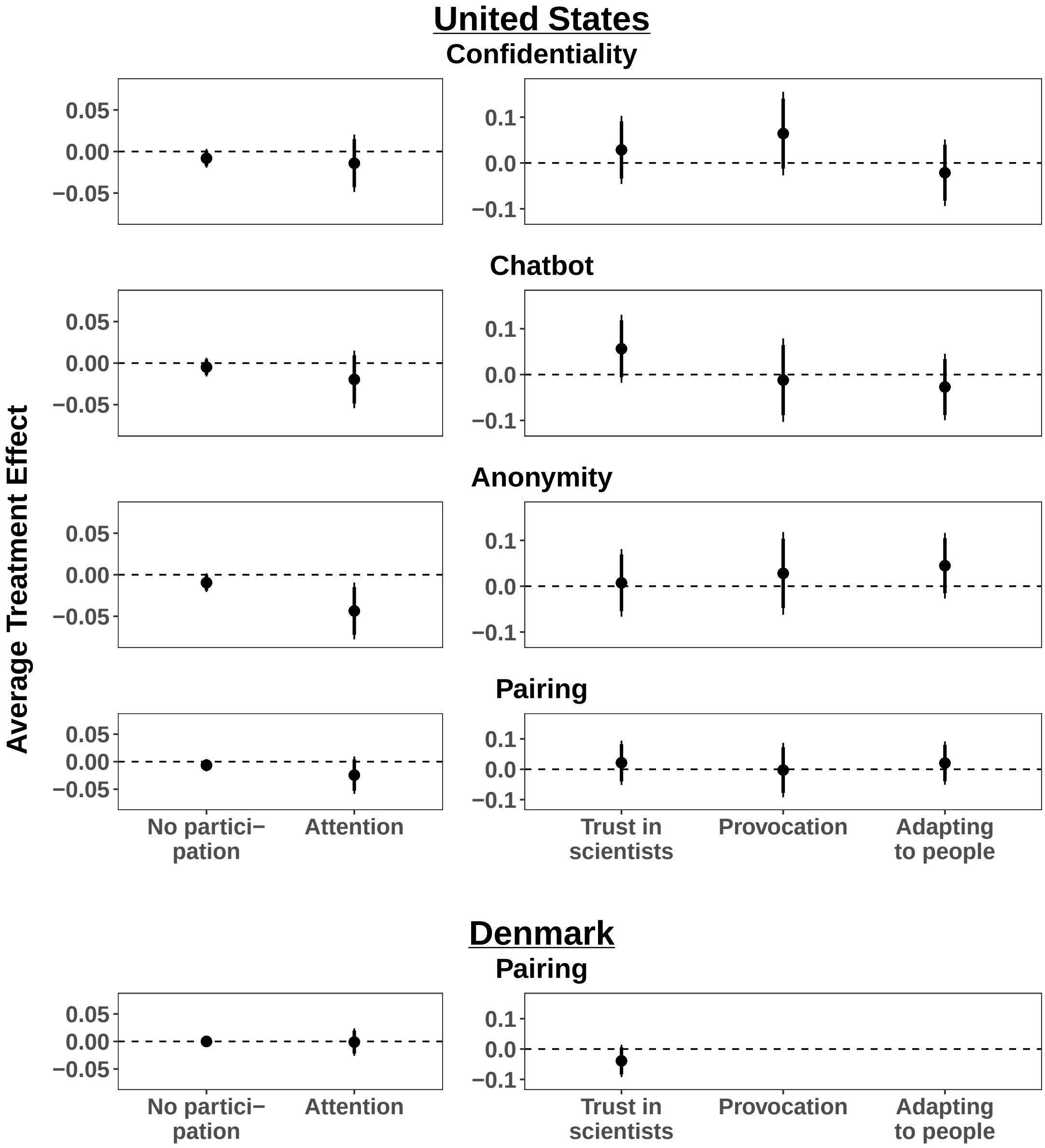
SI.2 Additional analyses

Figure SI 2: Treatment Effect on Emotions Felt by Respondents while Answering the Survey



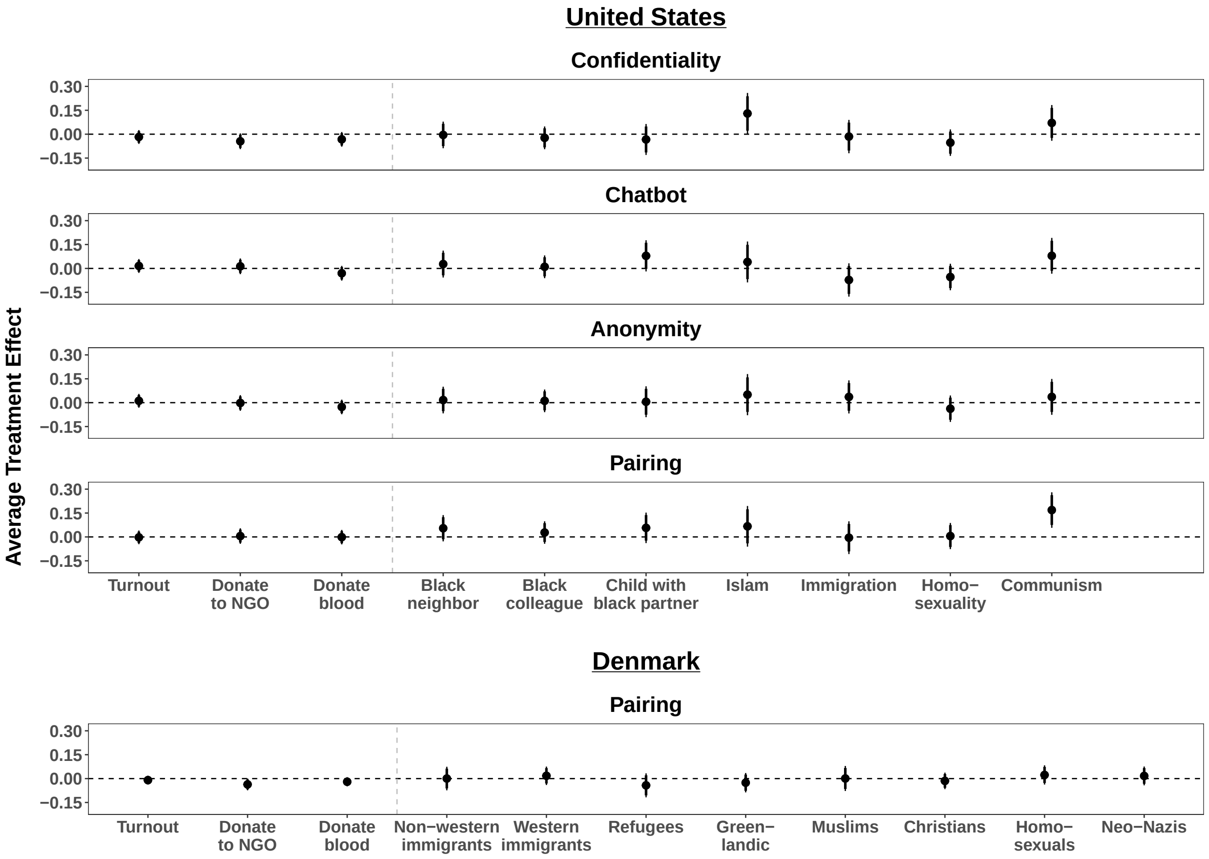
Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

Figure SI 3: Treatment Effect on Dropout, Non-participation, attention, and other outcomes



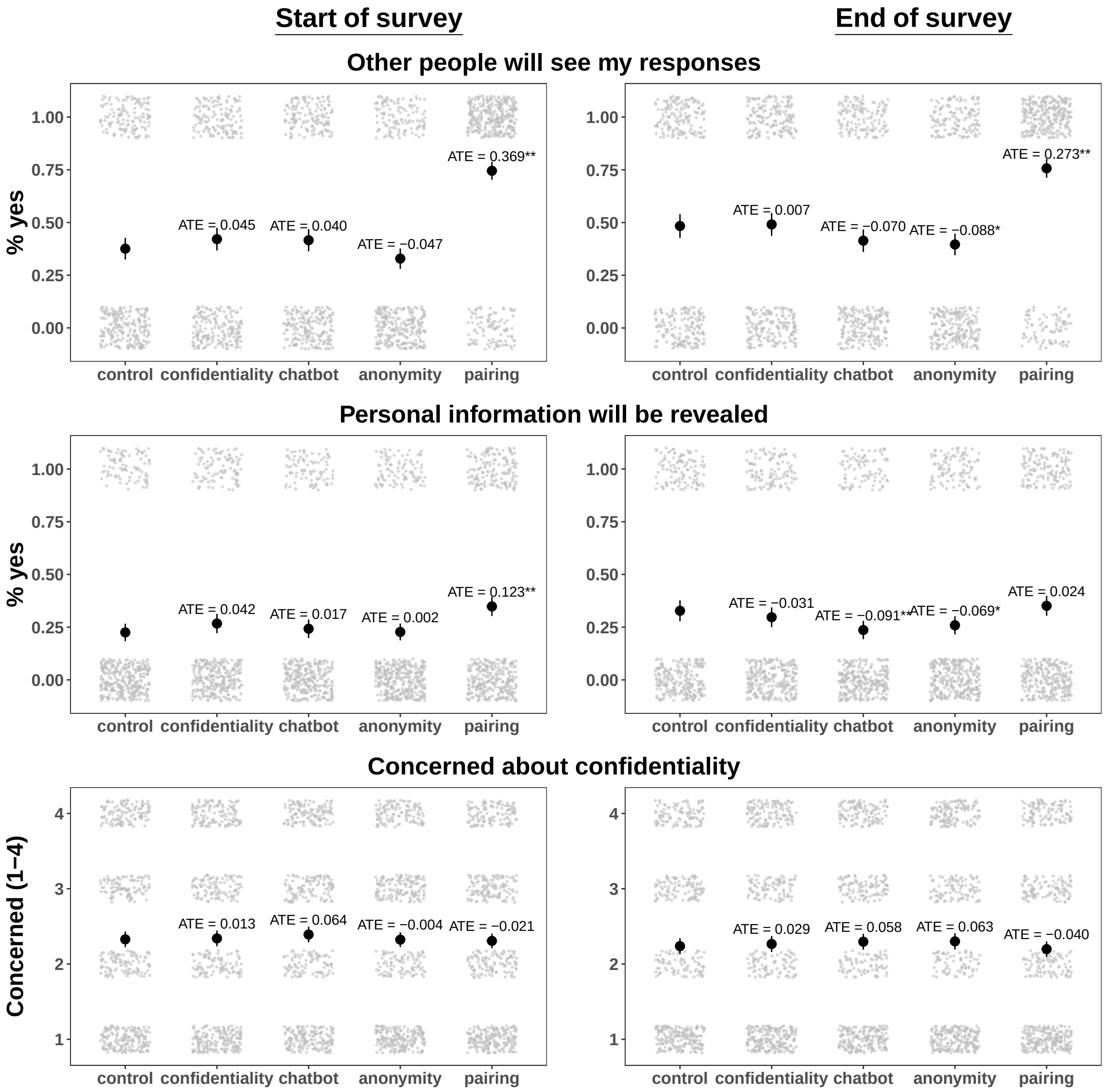
Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

Figure SI 4: Treatment effects among attentive respondents



Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

Figure SI 5: Manipulation depending on position of manipulation checks (United States)

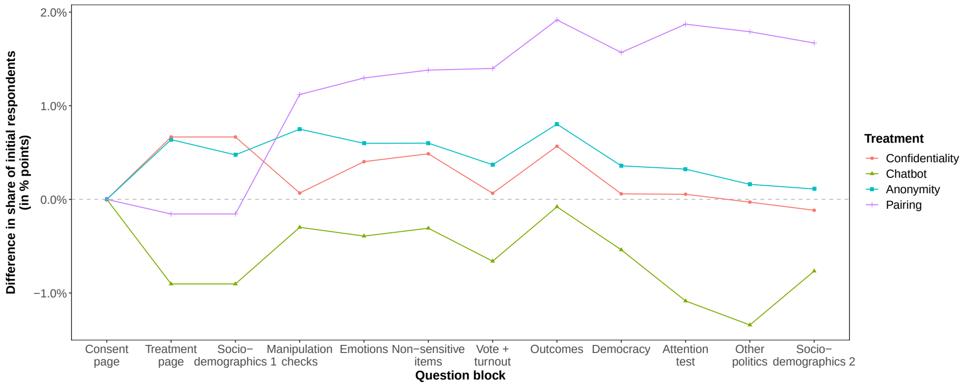


Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

SI.3 Dropouts during the survey

SI.3.1 United States

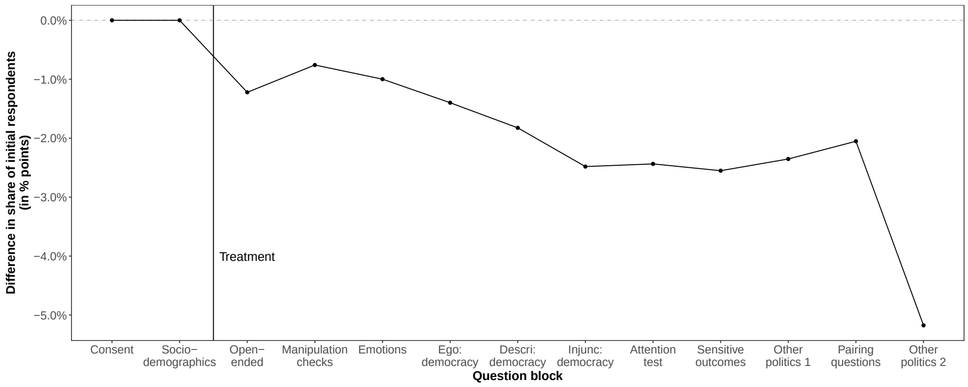
Figure SI 6: Dropout during the survey (United States)



Note: Reported are the % points of dropouts along the survey design.

SI.3.2 Denmark

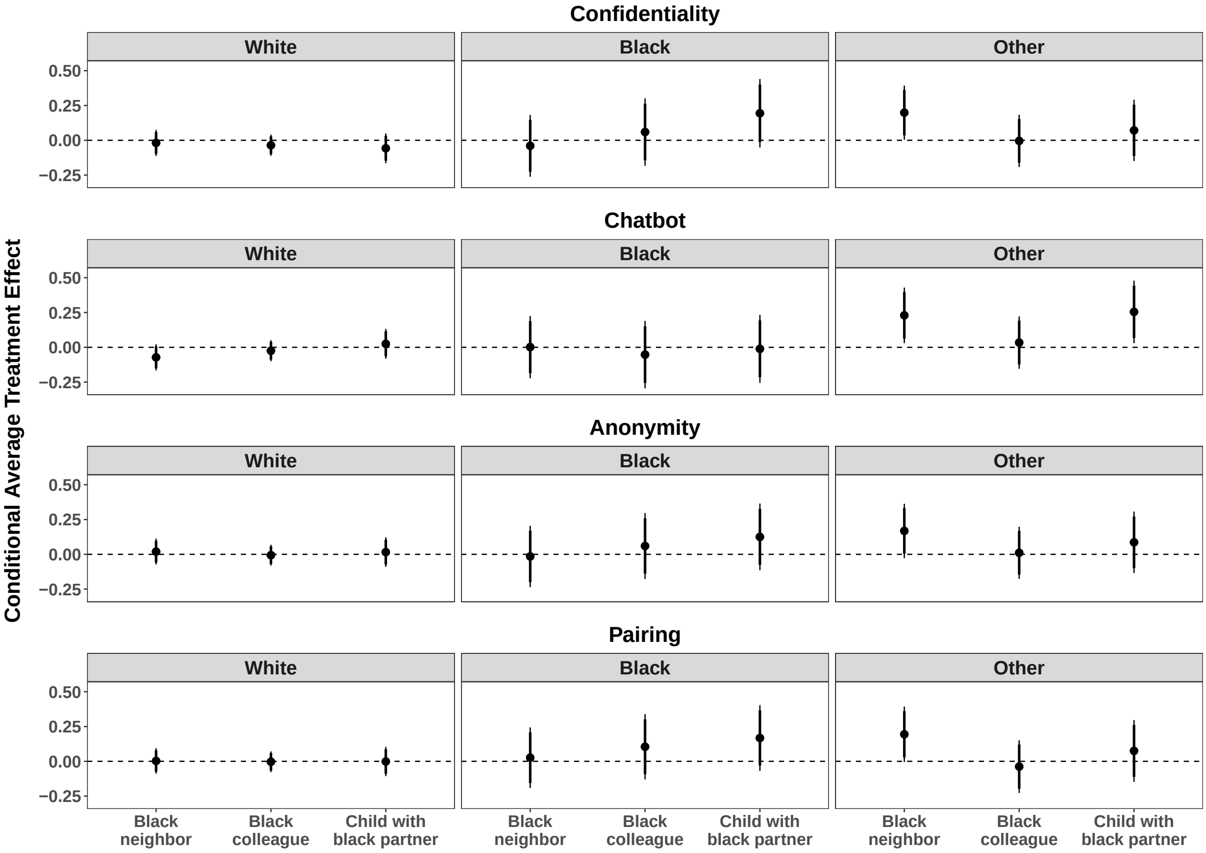
Figure SI 7: Dropout during the survey (Denmark)



Note: Reported are the % points of dropouts along the survey design.

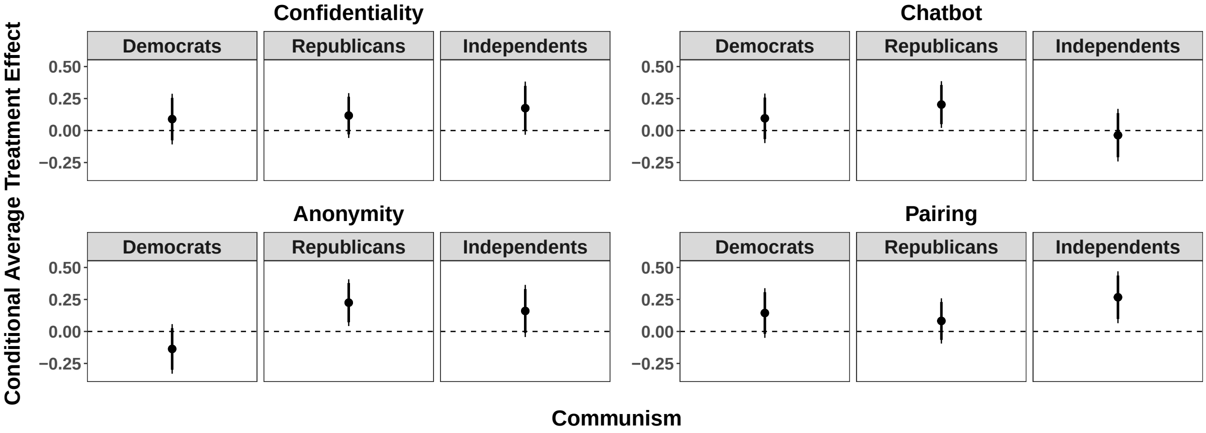
SI.4 Heterogeneous treatment effects

Figure SI 8: Heterogeneous treatment effects: racism



Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

Figure SI 9: Heterogeneous treatment effects: support for communism



Note: Reported are differences in means (point estimates) along with their 95% confidence intervals (whiskers).

SI.5 Our survey questions

SI.5.1 Treatment text

1. confidentiality: […] "The results from this survey, including your individual opinions, will be posted on our website in approximately one month after the analysis is completed. We will notify you when the results become available on our website (website link)" […]
2. anonymity: "Thank you for participating in our survey. Even though you indicated your willingness to participate on the former page, we want to re-emphasize that **your answers are completely anonymous and we treat your responses confidentially**"
   Have you read this information? Y/N
3. chatbot: pop-up with following text: "Hi! I am the research assistant of this project and I will be live with you while you respond to our survey. If any question emerges, please do not hesitate to get in touch by writing to me at the following address: morgan.lcj@ps.au.dk, and I will come back to you shortly."
4. pairing: "*Notice:* After you have finished we may show your responses to other respondents and you may also see responses from other respondents. These other respondents are fellow Americans responding to this survey – *just like you.*"
   Have you read this information? Y/N

SI.5.2 Manipulation checks

(While responding to our questions,) How much are (were) you concerned about the confidentiality of your responses?

1. not at all
2. a little
3. somewhat
4. a great deal

(While responding to our questions,) How much are (were) you concerned about other people seeing your individual responses?

1. not at all
2. a little
3. somewhat
4. a great deal

Do you think other people than the researchers will see your individual responses (for instance other respondents, fellow Americans)?

1. Yes
2. No
3. I don't know
4. I don't want to answer

Generally speaking, how do you feel while responding to our questions? Please tell us how much you feel each of the following emotions? [Angry, Disgusted, Enthusiastic, Happy, Anxious, Afraid]

1. Not at all
2. Not very much
3. Somewhat
4. A great deal

SI.5.3 Outcomes (United States:)

- **Turnout:** *Did you vote in the last Presidential election in 2020?*
- **Donations:** *Over the past 12 months, we would like to understand how frequently you engaged in specific behaviors. Please indicate how often you performed the following actions during this period.*
  - *Donated money to a non-profit organization such as UNICEF or the Red Cross.*
  - *Donated blood.*
- **Attitudes:** *To what extent do you agree or disagree with the following statements...?*
  - *I would like having a black family moving in next door.*
  - *I would feel comfortable if one of my colleagues at work was black.*
  - *I would feel comfortable if one of my children was in a love relationship with a black person.*
  - *Islam does not belong to our country.*
  - *Immigrants make this country a worse place to live.*
  - *Gays and lesbians should be free to live their own life as they wish.*
  - *Communism is better than it is commonly thought to be.*

SI.5.4 Outcomes (Denmark:)

- **Turnout:** *Did you vote in the 2022 general election?*
- **Donations:** *Over the past 12 months, we would like to understand how frequently you engaged in specific behaviors. Please indicate how often you performed the following actions during this period.*

  - *Donated money to a non-profit organization such as UNICEF or the Red Cross.*
  - *Donated blood.*

- **Attitudes:** *How would you feel about having one or more people from the following groups as a neighbor?*
  - *Non-western immigrants*
  - *Western immigrants*
  - *Refugees*
  - *Greenlandic*
  - *Muslims*
  - *Christians*
  - *Homosexuals*
  - *Neo-Nazis*

SI.6 Preregistration

Our anonymized PAP can be found under: https://osf.io/27ycb/?view_only=0fd0d1290e3a4821a7474e22277f662c

SI.6.1 Deviations from our pre-analysis plan

We made the following changes to the design of our survey after a soft launch of 157 respondents:

- The Supreme Court decided that alternative action is no longer allowed for college admission. This happened during our soft launch. We, thus, added another sensitive question about "having a black neighbor" to ensure that our results are not biased due to this decision by the court. **(we added a new sensitive question that we did not pre-register. This means that another question is analyzed without having been pre-registered.)**
- We added a question on trust in scientists to the trust battery. **(no effects on this paper)**

After the first launch of 1,000 respondents, several issues emerged which we then addressed:

- We added another manipulation test to the questionnaire. We did not find manipulation effects after the soft launch on any other treatment but the consent form. To make sure that these null findings are not driven by the wording of our original manipulation test, we added another one: "how much are you concerned about other people seeing your responses?" **(we analyzed this manipulation test in the results section of this paper.**
- We added another question after the soft launch on the need for social desirability after the soft launch: "Depending upon the people involved, I react to the same situation in different ways." We did this to rule out that the general tendency of respondents for being socially desirable is not affected by the treatments. **(we analyzed this manipulation test in the SI section of this paper.**
- We added another question to test for a "provocation" backlash effect of out treatment "I need chaos around me - it is too boring if nothing is going on."

After another 1,000 respondents still several issues emerged, among other things a high share of drop-outs on the "incentivize screen":

- We got rid of the incentives.
- Instead we moved some of the manipulation tests (questions on concern about others knowing) in before the political questions batteries to ensure that respondents are primed with the manipulations before answering the social norm question batteries.
- We added another manipulation test asking respondents "Do you think other people than the researchers will see your individual responses (for instance other respondents, fellow Americans)?" (Yes/No response categories). **(we analyzed this manipulation test in the results section of this paper.**

- We removed he wording “while responding to our questions” from our manipulation questions. The reason is that respondents do not need to think that they are observed while answering our survey. Instead, they should have the impression that eventually other human beings will see their individual responses.

Following the data collection and survey analysis of a 1,000 responses, we introduced manipulation check number one and two to address several issues.[7] First, we added both checks to get beyond an exclusive focus on confidentiality to include a more general focus on being seen by “other people”, which we believe is more representative of what we are trying to do. Second, we removed the emotional reference of being "concerned" to focus on a more neutral formulation, whether they thought their answer will simply be seen to disentangle our treatment effects on respondents’ perception of their answer being public with its effect on the emotional state of respondents.

[7] deviating from our pre-analysis plan, see : SI.6.1