When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments[†]

Graeme Blair[‡] Alexander Coppock[§] Margaret Moor[¶]

First draft: May 1, 2018
This draft: May 13, 2019

Abstract

Eliciting honest answers to sensitive questions is frustrated if subjects withhold the truth for fear that others will judge or punish them. The resulting bias is commonly referred to as social desirability bias, a subset of what we label sensitivity bias. The scope of sensitivity bias is important for researchers investigating social phenomena and also for governments seeking to understand public opinion. We make three contributions in this paper. First, we propose a social reference theory of sensitivity bias to structure expectations about survey responses on sensitive topics. Second, we describe the choice between direct and indirect measurement technologies as a bias-variance tradeoff. Third, to estimate the extent of sensitivity bias, we meta-analyze the set of published and unpublished list experiments conducted to date and compare the results with direct questions. We find that sensitivity biases are typically smaller than 10 percentage points and are in some domains approximately zero.

[†] Moor acknowledges the ISPS Dahl Scholarship for support. We thank Kosuke Imai and Rebecca Littman for many conversations on list experiments. We also thank Kirk Bansak, Jamie Druckman, Adam Glynn, Larry Hedges, Sophie Jacobson, Tyler Reny, and Aaron Rudkin, as well as audiences at the Q-Center and the Druckman Lab at Northwestern University and APSA for helpful comments.

[‡] Graeme Blair is Assistant Professor of Political Science, University of California, Los Angeles. `https://graemeblair.com`

[§] Alexander Coppock is Assistant Professor of Political Science, Yale University. `https://alexandercoppock.com`

[¶] Margaret Moor is a graduate of Yale College, Class of 2018.

Scientific survey research traces its origins to George Gallup's first nationwide sample survey in 1936. Since their introduction, researchers have worried that survey responses suffer from misreporting and nonresponse due to the sensitivity of some questions (Maccoby and Maccoby 1954). In a small number of well-documented cases, validation studies have demonstrated that survey estimates of sensitive traits may be distorted. For example, the voter turnout rate in the U.S. is below 60%, but survey estimates put it at between 75 and 85% (Ansolabehere and Hersh 2012). In the other direction, one meta-analysis found that 30-70% of clinically-confirmed recent drug users reported they had not used drugs (Tourangeau and Yan 2007).

We call this form of measurement error sensitivity bias. The vast extant literature on misreporting and nonresponse in sensitive settings often invokes the term "social desirability bias." In our view, that term is imprecise. First, it conflates the sensitivity of the topic with the properties of the measurement tool. Second, it leaves open to interpretation "who" desires a particular response and why a respondent would care. In this paper, we build on frameworks from social psychology and political science to advance a social reference theory of sensitivity bias that disentangles these considerations.

Cottage industries have emerged in nearly every social science discipline to address sensitivity bias. Techniques fall into three broad categories: changing the form of the question (Haire 1950; Warner 1965; Miller 1984), changing the context of how the question is answered (Silver et al. 1986), and identifying which types of people are most prone to giving false answers (Snyder 1987; Paulhus 1991; Berinsky 2004). Each approach comes with costs, in terms of development and testing, survey duration, and statistical power. Despite 70 years of methodological innovation, it remains difficult for researchers to decide whether sensitivity bias is a problem and if it is, how best to address it.

In this paper, we tackle these questions in three parts. First, we introduce a theory of sensitivity bias to structure thinking about whether bias is likely. Applying our theory to a given empirical setting requires articulating a social referent in particular (for example, the self, a spouse, neighbors, or the state), respondents' perceptions of the likelihood that responses will be revealed to that individual or group, and the perceived consequences of the revelations.

Second the choice among measurement technologies to address sensitivity bias amounts to a bias-variance tradeoff. Direct questions may be biased but they are low variance. Alternative question formats, such as the list experiment (Miller 1984), the randomized response technique (Warner 1965), or the cross-wise technique (Gingerich et al. 2016) may exhibit less bias but are higher variance. Because the list experiment is by far the sensitive question format of choice among political scientists, we restrict our discussion of the bias-variance tradeoffs to the choice between direct questions and list experiments, but the broad strokes of our argument apply to these other formats as well.

Third, we compare 30 years' worth of list experiments with direct questions in order to estimate the extent of sensitivity bias in many substantive domains. Our census of published and unpublished list experiments extends from the first list experiment published in 1984 up through the end of 2017, when we concluded data collection. This body of research covers topics of most interest to political scientists such as racial prejudice, turnout, and vote buying, but also criminal behavior, sexual activity, and drug and alcohol use. Our results indicate that sensitivity bias is typically small to moderate, *contra* the evident expectation on either the authors' or their real or imagined reviewers' parts that misreporting was a large concern. However, there is considerably heterogeneity in sensitivity bias across subject domains. We find evidence of overreporting voter turnout, underreporting vote buying, but nearly no evidence of sensitivity bias in measures of prejudice.

Determining whether sensitivity bias is a problem in a particular domain is often a matter of intuition, conjecture, or previous theoretical expectations. Researchers can use our empirical results to recalibrate their expectations of sensitivity bias and to reassess their position on the bias-variance frontier. At typical levels of sensitivity bias, list experiments are far less accurate (in terms of mean-squared error) than direct questions when sample sizes are smaller than 3,000 subjects.

Theories of Sensitive Survey Responses

Why do questions about sensitive topics in surveys such as drug use and voter turnout generate biased responses? We develop a social reference theory of sensitivity bias that distinguishes between the sensitivity of the topic and the properties of the measurement tool (typically self-reported responses to direct questions in sample surveys).

Defining Sensitivity Bias

To formalize sensitivity bias, imagine that a subject i harbors a latent true value of the sensitive trait D_i^*. We distinguish the latent value from the response that a subject would give if asked directly, D_i. We assume that subjects know D_i^* and that the survey question is designed with sufficient construct validity that subjects understand that researchers are asking subjects to report D_i^*. If the subject does not respond when asked, D_i is missing (i.e. $D_i = \texttt{NA}$). In the most general terms, measurement error occurs if $D_i \neq D_i^*$, i.e., if there is any slippage between the latent trait and the survey response. Measurement error may result from many different causes, including technical slip-ups, miscommunications between respondent and interviewer, or even deliberate falsification of responses by survey staff. We are concerned here with the special form of measurement error that occurs when $D_i \neq D_i^*$ because of the sensitivity of the question. A common target of inference is the average value of D_i^*, or the prevalence rate $\pi^* \equiv \mathbb{E}[D_i^*]$. If survey reports are distorted by sensitivity bias, then direct questions may only estimate $\pi \equiv \mathbb{E}[D_i \mid D_i \neq \texttt{NA}]$, which equals the prevalence rate plus a bias term. Sensitivity bias may be defined as $\delta \equiv \pi - \pi^*$. If δ is positive, direct questions exhibit overreporting and if δ is negative, they exhibit underreporting.

Our model of sensitivity bias requires that a (unique) latent value D_i^* exists for each subject. This assumption would be violated if subjects do not harbor specific attitudes and beliefs (even if they would respond when asked a question on a survey). Further, we do not consider settings with "multiple truths," which would mean that D_i^* is random or vector-valued. In order to speak coherently about sensitivity bias, we have to imagine there is a true, scalar latent trait D_i^* that may or may not be different from the survey response D_i.

A Social Reference Theory of Sensitivity Bias

The dominant explanation for sensitivity bias since the 1950s has been social desirability bias (Maccoby and Maccoby 1954). According to Fisher (1993, pp. 303), social desirability bias results from "the desire of respondents to avoid embarrassment and project a favorable image to others." Goffman's *The Presentation of the Self in Everyday Life* (1959) launched research inquiries across sociology and social psychology into the importance of impression management or self-presentation (for a review, see Leary and Kowalski 1990). Goffman argues that people have in their own minds an idea of how they are perceived by others and take actions to improve that perception. Social desirability bias is a behavioral manifestation of self-presentation. Beyond social desirability, scholars have identified self-image, the fear of disclosure of responses, and intrusive topics as additional causes of sensitivity bias.

Three elements of a survey jointly determine if an item will be affected by these biases: the topic of the question (is it sensitive or not), the format of the question (is the question asked directly and what assurances of anonymity are made), and the context in which it is asked (who is listening to responses, and who can read or hear them after the interview).

The last element highlights the fact that we must know *with respect to whom* respondents manage impressions. Psychologists and political scientists have developed and applied scales to measure person-constant levels of desirability bias (Snyder 1987; Paulhus 1991; Berinsky 2004). Yet clearly the set of actors who the respondent believes can hear or read responses, during and after the interview, as well as other elements of the survey context may influence the strength of impression management pressures and the risk of disclosure. We interpret the evidence showing that different people withhold at different rates as a consequence of individuals' idiosyncratic beliefs about the relevant social referent. Respondents hold beliefs about who is asking questions, who sent the interviewers to ask, who can overhear the responses, and who can read responses after the interview is conducted.

Beliefs may be heterogeneous across contexts and across respondents. For example, we demonstrate in the supplementary materials that respondents to the Afrobarometer vary widely in their perceptions of the survey sponsor (see also Corstange 2014).

Perhaps the most salient social referent for subjects is the interviewer asking the survey question (Feldman et al. 1951; Simpser 2017). Subjects may presuppose (rightly or wrongly) that survey-takers have an opinion about what the correct attitude to hold is. Using randomized experiments of interviewers to respondents, interviewer effects have been demonstrated for interviewer race (Hatchett and Schuman 1975; Cotter et al. 1982; Davis 1997), gender (Kane and Macaulay 1993; Catania et al. 1996; Huddy et al. 1997), and perceived religiosity (Blaydes and Gillum 2013). Bystanders, family members, coworkers, or others who may be within earshot constitute different sets of social referents (Silver et al. 1986). Subjects might feel constrained to respond in a particular manner or not at all if under the watchful eye of a spouse (Aquilino 1993). Other more distant social referents may include those who will read responses after the survey ends, such as the sponsoring institution or academic analysts, consumers of the survey data including the media and the general public, or more worryingly, the government or armed groups who might take punitive action depending on the response.

Social desirability is not the only source of sensitivity bias. First, respondents face pressures that come from themselves, not only others (Greenwald and Breckler 1985). Second, questions may be seen as "intrusive," representing taboo topics respondents feel are out-of-bounds independent of perceived social desirability (Tourangeau et al. 2000). For taboo topics, the act of responding, separate from the content of the response, may itself be sensitive. In this case, sensitivity bias will manifest as nonresponse. Third, respondents may fear their responses will be disclosed to authorities such as governments, criminals, armed groups, or employers. In such cases, the worry is not what is socially desirable, but instead what is safe.

We synthesize these strands into a social reference theory of sensitivity bias. Sensitivity bias occurs when all four of the following elements are present:

1. A social referent (one or more people or organizations) the respondent has in mind when considering how to respond to a survey question. A social referent could be the respondent him or herself.

2. A respondent perception that the social referent can learn the subject's response to the sensitive question.

3. A respondent perception of a descriptive or prescriptive social norm about what response (or nonresponse) the social referent desires.

4. A respondent perception that failing to provide the response desired by the social referent would entail costs to themselves, other individuals, or groups. Costs may be social (embarrassment), monetary (fines), or physical (jail time or personal violence).

Before applying this theory to some common political science research settings, a few brief comments. First, the same question may be sensitive in one context but not another, depending on the social referents that may be relevant in each. Respondents may perceive that different referents prefer different answers, which raises the possibility of cross-cutting sources of bias. Second, this theory helps distinguish sensitivity bias from other forms of measurement error. In particular, we draw a fine distinction between self-deception and recall failures. If respondents misreport because they do not want to admit, even to themselves, that they participate in the sensitive behavior, direct questions will suffer from sensitivity bias. If however, respondents simply do not spend sufficient cognitive energy to recall whether, for example, they voted in the most recent midterm election, direct questions will be biased, but not because of sensitivity. Third, we often think of sensitivity bias in terms of its effects on self-reported attitudes, but it may apply equally to survey reports of behaviors. Whether sensitivity bias differs between attitudes and behaviors is an empirical question that to our knowledge has not been investigated. Measures of both could be distorted by sensitivity bias, depending on the social referent, the perceived risk of disclosure to that social referent, perceived social norms, and the perceived costs. Finally, a voluminous research literature examines the distinction between implicit and explicit attitudes (Greenwald and Banaji 1995; Greenwald et al. 1998). Since implicit attitudes are thought to operate at an unconscious level, respondents themselves are likely unable to accurately self-report them. We only consider sensitivity bias that may plague measures of explicit attitudes (Littman 2015).

Sources of Sensitivity Bias in Four Political Science Literatures

In this section, we apply our theory to four political science research literatures in which sensitivity bias has been flagged by researchers as a major source of measurement error. We reinterpret their reasoning through the lens of the four criteria for sensitivity bias. We present a (not necessarily exhaustive) list of social referents identified in each literature and how the criteria are applied in each literature in Table 1.

Clientelism in developing countries

The dominant mode of politics in many developing countries is clientelism, in which material goods are exchanged by politicians and voters in return for votes on an individual basis rather than on the basis of need as in programmatic political systems (for a review of accounts in political science, see Stokes 2007). Despite considerable scholarship on how clientelism works, we know relatively little about how pervasive it is, and whom it targets (Weitz-Shapiro 2012). Yet vote buying, the main behavior that characterizes clientelistic systems, is by its nature hidden. As a result, survey research – asking voters if they exchanged their vote for goods – is required both to measure its prevalence and whom is targeted by offers of exchange.

A recent flurry of scholarship has probed whether survey measures of vote buying are distorted by sensitivity bias. Vote buying is illegal in most places, so respondents may have a reasonable fear of prosecution (Mexico, Imai et al. (2014); Lebanon, Corstange (2017); Singapore, Ostwald and Riambau (2017); Nicaragua, Gonzalez-Ocantos et al. (2012); Hungary, Mares and Young (2016)). In some contexts, however, voters may not be concerned about the illegality of vote buying because of lax enforcement. For example, in a study of vote buying in the Philippines, Cruz (2019) speculates

	Respondent beliefs		
Social referent	Referent can obtain answer	Referent's preferred answer	Cost if preferred answer not provided
Clientelism: "Did you exchange your vote for money, gifts, or services?"			
Interviewer	Yes, provided directly	No	Self presentation
State authorities	Possibly, depending on anonymity protections	No	Arrest
Neighbors	Possibly, depending on anonymity protections	No	Shame
Politician(s) who exchanged vote	Possibly, depending on anonymity protections	No	Will not offer exchange in future
Prejudice: "Would you feel angry or upset if a black family moved in next door to you?"			
Interviewer	Yes, provided directly	No	Self presentation
Self	Yes	No	Self image
Support for authoritarian regimes: "I voted for Vladimir Putin in the most recent Presidential elections."			
State authorities	Possibly, depending on anonymity protections	Yes	Arrest
Voter turnout: "In the presidential election of November 8, 2016, did you vote?"			
Interviewer	Yes	Yes	Self-presentation
Household members within earshot	Possibly, depending on anonymity protections	Yes	Self presentation
Self	Yes	Yes	Self image

Table 1: Possible sources of sensitivity bias in four political science literatures. An example of the sensitive survey item is given for each research literature.

that the low levels of sensitivity bias in direct questions about the practice may be explained by the high prevalence of vote buying or because, "laws that forbid it are rarely enforced in the Philippines" (pg. 390). Respondents may be reluctant to admit selling their vote because of "the implication that they are poor enough to sell their votes" (Stokes 2005, pg. 321). Similar logics have been forwarded for Lebanese (Corstange 2017) and Nicaraguan (González-Ocantos et al. 2015) respondents. Beyond the possible association with low socio-economic standing, respondents may wish to avoid being seen as a participant in an immoral or unethical behavior (Bulgaria and Romania, Mares et al. (2016)) or to "acknowledge that the handout influenced their vote" (Brusco et al. 2004, pg. 69). Respondents may also wish to avoid appearing to have violated perceived social norms about behavior as a democratic citizen (Kramon 2016). Finally, Frye et al. (2014) highlights in the Russian case that respondents may have in mind their employer as a social referent, who may have special levers of retaliation.

Prejudice

The frequency and intensity of outright expressions of racial prejudice towards black Americans by white Americans has declined over time, but the causes and consequences of this change remain sources of scholarly and public debate about racial attitudes (Bobo 1988; Schuman et al. 1997). A central theme of the debate is whether old-fashioned racism has been supplanted by a modern form of racism or if little has changed but whites' willingness to express their racist views in public (Kinder and Sears 1981; Tesler 2012). The survey research in this literature is beset by deep measurement difficulties, including disagreements about what the releveant theoretical constructs are and what survey questions might measure them (Sniderman and Tetlock 1986). One point of agreement, however, is that sensitivity bias could undermine any of the measures if respondents believe that interviewers prefer a particular answer and would judge the respondent to be a racist if that answer were not given. For this reason, the prediction is that, if anything, respondents underreport racist attitudes. The measurement problem is compounded by the difficulty (or impossibility) of separating attitudes towards policies like welfare or affirmative action from racial attitudes (Gilens 2009). However, if *respondents* think interviewers think those policy views are the result of racist attitudes, sensitivity bias could infect measures of policy attitudes regardless of the true causes of policy views.

The usual logic of sensitivity bias for racial attitudes extends directly to possible biases in measures of attitudes in other domains such as sexual orientation and religion. Respondents may wish to avoid being seen as retrograde or bigoted, so they may overreport positive attitudes and underreport negative attitudes. Especially in view of the dramatic shifts in public opinion on same-sex marriage, it is reasonable to wonder whether some or all of these changes can be attributed to sensitivity bias.

Similarly, religious tolerance and respect for members of other faiths is a widely expressed value in many cultures. The evident social and cultural divisions along religious lines raise the possibility that survey affirmations of religious tolerance are disingenuous.

Support for authoritarian regimes

At the heart of many models of authoritarian politics is the need for dictators to assess public opinion in order to predict and prevent revolution. This may explain why dictators often hold elections and tolerate limited protests (Gandhi and Lust-Okar 2009). Dictators also rely on direct questions in surveys. In each case, the regime faces what Wintrobe (2000) calls the "dictator's dilemma," in which the regime needs to know its true support to maintain stability, but publicly revealing dissatisfaction may itself lead to instability. As a result, dictators may exert pressure on citizens to profess higher levels of regime-support opinions than they truly hold (a phenomenon labeled as "preference falsification," see Kuran 1997) and prevent polls that reveal low levels of support from being conducted or published. A casual inspection of recent polls on leader approval suggest this is exactly what is happening: as many as 89 percent of Russians report support for Russian President Vladimir Putin in high-quality surveys from Russia's only independent polling agency (Yuri Levada Analytical Center 2019) and 93 percent reported trust in the central government in the 2012-13 World Values Survey China (Tang 2016). Indeed, scholars of authoritarian regimes argue the four sensitivity bias criteria are plausibly met: (1) the regime is a top-of-mind social referent when answering questions; (2) regime informants can plausibly uncover responses to surveys; (3) citizens know the responses the regime prefers, learned through propaganda; and (4) costs may include harassment, imprisonment, or worse (Frye et al. 2017; Chen and Yang 2018).

The stakes to learning truthful answers are high for the regime, who wishes to stay in power and take actions to placate potential revolutionaries; for opposition supporters, who may have the false impression that regime support is high may feel they cannot risk going into the streets; and for scholars and foreign intelligence analysts in the business of predicting and mitigating the consequences of revolution. Political scientists failed to predict the fall of the Berlin wall, and in the aftermath focus turned squarely to the disjuncture between "public truths and private lies" (Kuran 1991). In some (posthoc) accounts, revolution only happened after an information cascade following small protests in Leipzig, which revealed to East Germans their shared antipathy to the regime and willingness to act for change (Lohmann 1994).

Voter turnout

From the earliest investigations into voter behavior in America, scholars have documented that that estimates of voter turnout based on survey self-reports are upwardly biased. Campbell et al. (1960, pp. 93-96) report that 74% of the first National Election Study (1952) sample reported voting, whereas the official turnout rate in that election was only 63% (see also Burden 2000; Vavreck 2007). The possible sources of error between the survey prevalence rate and the true turnout rate are many: survey nonresponse, item nonresponse, or misreporting. Distinguishing between these sources of error was frustrated by the difficultly of matching survey respondents to individual voter file records. Some of these technical challenges have been overcome and the most recent studies have concluded that misreporting is a major contributor to the discrepancy (Ansolabehere and Hersh 2012; Enamorado and Imai 2018). Misreporting itself may or may not be due to sensitivity bias as we have conceived of it here. Holbrook and Krosnick (2010) subdivides misreporting into "memory errors" and "social desirability bias." Memory errors occur when respondents do not invest the cognitive effort to be sure they did or did not vote in the particular election the interviewer is asking about. It is possible that respondents remember having voted in some past election and so are more likely to respond yes when asked if they participated in a particular election. Belli et al. (1999) show that some direct question wording variants are able to reduce memory errors of this sort. The list experiment is *not* designed to help with reducing memory errors, but may give subjects cover to admit to survey takers that they did not vote. Because whether or not a person has voted is public, subjects are unlikely to be specifically worried about state authorities discovering their survey response. Therefore, the social referents that a survey subject must have in mind are the interviewers, household members within earshot of the interview, and themselves. In all cases, we imagine that the perceived cost of the social referent learning (or re-learning, in the case of the "self" social referent) is shame at having failed in a civic duty.

List experiments to reduce sensitivity bias

The list experiment, also known as the item count technique and the unmatched count technique, hides individual responses to a binary sensitive item by aggregating them with the answers to several binary control items.[1] Sensitive item responses are hidden from interviewers, bystanders, and data consumers who only learn the count including control items. The prevalence rate of the sensitive item can be estimated by subtracting the average count of the control items from the observed count, leaving only the sensitive item. The list experiment is designed to mitigate sensitivity bias by minimizing self-presentation concerns[2] and the risk of disclosure. Self-presentation concerns of the respondent are addressed by hiding the sensitive item response from the interviewer, bystanders, or later data consumers. Presentation concerns about self-image are not addressed, however. The list experiment also minimizes the risk of disclosure. Authorities, such as employers or parents, cannot exactly identify the sensitive item response for most respondents, except in the case of floor and ceiling responses (Glynn 2013). The list experiment does not change the bias from asking intrusive questions, because the text of the question includes the same sensitive item text found in the direct question.

We illustrate the logic of the list experiment with an example. Kramon (2016) reports on a nationally-representative survey of 1,927 Kenyans administered after the 2007 Kenyan elections. The survey focuses on estimating the proportion of voters who experienced vote buying attempts

[1] We present a formalization of the list experiment and recapitulate the list experiment assumptions of "No liars" and "No design effects" in the supplementary materials.

[2] For a formalization of a related argument focusing on respondent second-order beliefs about the interviewer, in terms of prior beliefs about the respondent's sensitive trait and posterior beliefs following learning the response to the list experiment, see Simpser (2017).

during the election. To do so, the authors use a list experiment and a direct question. Respondents were randomized into three groups: a control group and two treatment groups. In the control group, respondents were read the following instructions:

```
Election campaigns are a busy time in our country.  I am going to read
you a list of some of things that people have told us happened to them
during the 2007 campaign.  I am going to read you the whole list, and
then I want you to tell me how many of the different things happened
to you.  Please do not tell me which of the things happened to you,
just how many.  If you would like me to repeat the list, I will do so.

1.  Politicians put up posters or signs in the area where you live.

2.  You read the newspaper almost every day to learn about the campaign.

3.  You met a politician personally to discuss his or her candidacy.

4.  You discussed the campaign with a chief or another traditional leader.
```

Responses in this control group represent the total number of control items to which the respondent answered "yes." In the "Influenced" treatment group, the same script was read but with a fifth item added to the list:[3]

```
5.  You voted for a party or politician because they gave you money
during the campaign.
```

In the "Received" treatment group, the fifth item read:

```
5.  You received money from a party or politician.
```

Thus, in the treatment groups, responses represent the count of the number of "yes" responses to the set of control items and the sensitive item. The observed data for treatment and control are displayed in Table 2.

Using data from the Kramon (2016) postelection survey in Kenya, we estimate the prevalence rate of vote buying, the main quantity of interest in the study. Figure 1 presents the results. The "influence" question appears to be affected by sensitivity bias. The list experiment estimate, while

[3] Typically, item order is randomized and the sensitive item is not necessarily the last item.

Count	Control	"Received" Treatment	"Influenced" Treatment
1	290	235	215
2	235	280	204
3	72	96	113
4	25	30	29
5		12	8

Table 2: Observed list experiment responses by treatment status for whether a bribe was received (third column) and whether the bribe influenced the respondent's vote (second column) from the 2007 Kenya postelection survey reported in Kramon (2016).

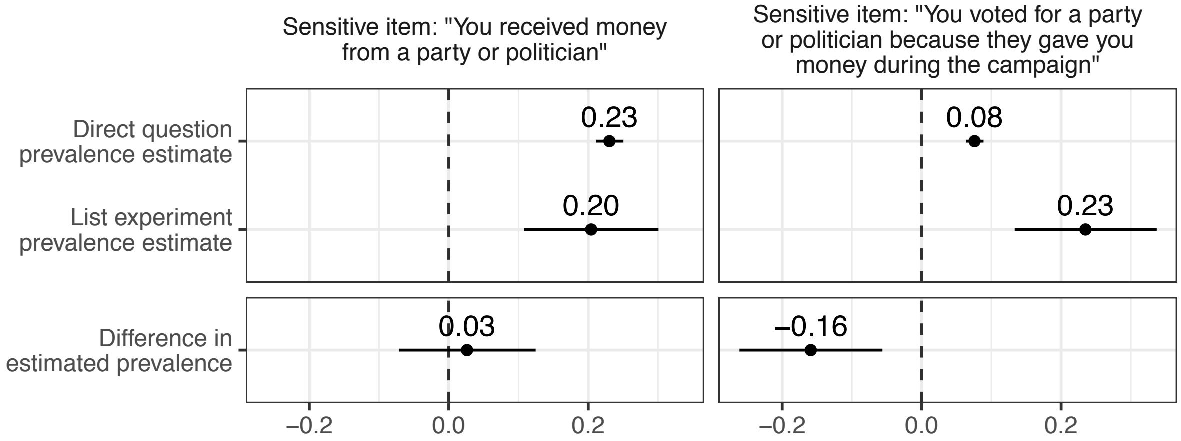


Figure 1: Estimated prevalence of vote buying from the list experiments and the direct question for two sensitive items presented in Kramon (2016): whether the respondent "received" a bribe, and whether a bribe "influenced" the respondent's vote.

imprecisely estimated, is definitively higher than the direct question estimate. By contrast, the direct and list experiment estimates of the proportion of respondents who received money from parties or politicians are quite similar.

Tradeoffs in the Choice of a Measurement Design

The choice between list experiments and direct questions presents a bias-variance tradeoff. Direct questions may be biased, but they produce low-variance estimates.[4] Given their identifying assump-

[4] In some cases, the sign and plausible magnitude of the bias may be known, which could enable researchers to recalibrate their prevalence estimates.

tions, list experiments are unbiased, but high variance.

Consider a study of $n = 2{,}000$ subjects sampled from a larger population with a true prevalence rate (π^*) of 50%, but that has a sensitivity bias (δ) of the direct question of 10 percentage points. D_i is the response that subject i gives to the direct question. The direct question estimator $\widehat{\pi}$ is the sample mean, $\widehat{\pi} = \frac{1}{n}\sum_1^n D_i$. The variance of the direct question estimator ($\widehat{\pi}$) is given by

$$\mathbb{V}(\widehat{\pi}) = \frac{\pi^*(1-\pi^*) + \delta(1-\delta) + 2(\delta - \pi^*\delta)}{n-1}. \tag{1}$$

The variance of the list experiment estimator $\widehat{\pi^*}$ is given by

$$\mathbb{V}(\widehat{\pi^*}) = \frac{4\mathbb{V}(Y_i(0) + \pi^*(1-\pi^*)}{n-1}, \tag{2}$$

where $\mathbb{V}(Y_i(0))$ is the variance of the control item response. See the supplementary materials for derivations of these variance expressions.

Plugging in the assumed values of $n = 2{,}000$, $\pi^* = 0.5$, and $\delta = 0.1$ and taking the square root yields a standard error of the direct question estimator of 0.015, or 1.5 percentage points. Assuming the variance is equal to 0.75 (the average control group variance in the set of list experiments we describe below), we obtain a standard error for the list experiment of 0.04, or 4 percentage points. For the same number of subjects, the list experiment is $(4/1.5)^2 \approx 7$ times more variable than the direct question. Stated differently, a researcher would need a sample of 14,000 subjects in order to produce a list experiment estimate as precise as the direct question with 2,000. The intuition for this stark shortcoming of the list experiment is that only half the sample is asked about the sensitive trait and their responses are further obscured by adding noise (the control count). In the supplementary materials, we discuss innovations in the design of list experiments that aim to mitigate this difficulty.

This bias-variance tradeoff interacts with the goal of the research. We identify three main goals: estimating a prevalence rate, demonstrating the presence of sensitivity bias, and estimating the *dif-*

ference in prevalence rates across groups.[5]

When the primary research goal is obtaining a good estimate of the prevalence rate of a sensitive trait (as in Gervais and Najle (2018), which sought to estimate the proportion of U.S. residents who identify as atheist), it is unclear whether the direct question or the list experiment will render estimates that are closer to the true prevalence rate in terms of root mean squared error (RMSE). The main parameters that govern which approach will be closer are the extent of bias and the sample size of the study. The left panel of Figure 2 provides a visual explanation of how these factors interact. All else equal, the higher the true bias of the direct question, the more we prefer the list experiment. However, for many sample sizes, the direct question has lower RMSE, even in the face of substantial sensitivity bias. The line in the figure describes the bias/sample size combination at which researchers should be indifferent between the two methods on the basis of RMSE. For a study with 1,000 subjects, the bias must be greater than 6 points to prefer a list experiment; at 2,000, the bias must be greater than 4.5 points. Figure 2 is generated assuming a true prevalence rate of 0.5, but because of the relatively small influence of the true prevalence rate on the variance of the list experiment, the results are quite similar regardless of prevalence rate.

Another goal in some settings is to show that a particular domain is or is not plagued by sensitivity bias by conducting both a direct question and a list experiment and comparing the results. For example Lax et al. (2016) find that the estimated difference in support for same-sex marriage is not statistically significant. The design parameters that govern the power of this procedure to detect sensitivity bias are again the true level of bias and the sample size. The left panel of Figure 2 plots the bias / sample size combinations at which the power to detect sensitivity bias is 80%. At 1,000 subjects, the bias would need to be 20 percentage points in order to reach 80% power; even at a sample size of 2,000, power to detect biases of 10 percentage points is well below the conventional power target.

[5] Other uses of sensitive questioning techniques include using the predicted values of the sensitive items as predictors in a regression, as in Imai et al. (2014).

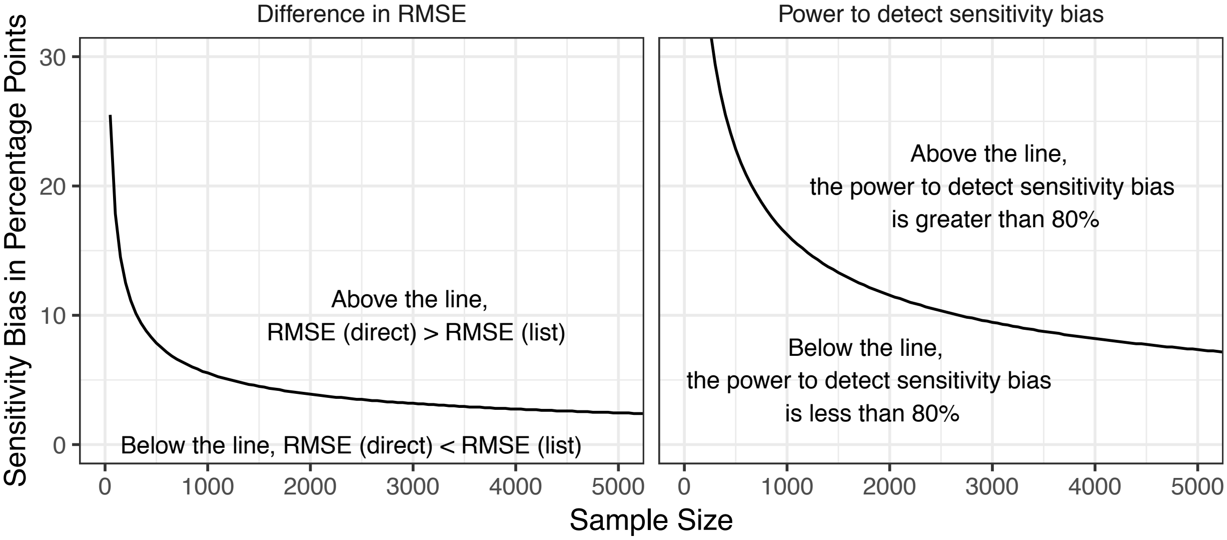


Figure 2: For a Given Sample Size, Whether the List Experiment is Preferred to Direct Questions Depends on the Expected Level of Sensitivity Bias. For designs that aim to estimate the prevalence rate of the sensitive item (left panel), the root mean-squared error indifference curve characterizes the tradeoff. When the goal is to detect sensitivity bias, the power indifference curve (right panel) indicates the level of sensitivity bias that suggests choosing the list experiment over direct questions.

In the supplementary materials, we show that the power and RMSE considerations for the third goal, estimating or demonstrating a difference in prevalence rates are even more daunting.

Meta-Analysis Research Design

In this section, we present a meta-analysis of list experiments to characterize the level of sensitivity bias in six political science literatures: turnout, prejudice on the basis of race, religion, and sexual orientation, vote buying, and political attitudes in authoritarian contexts. To do so, we compare responses to questions asked directly and in a list experiment. Within each research literature, we present the empirical distribution of estimated bias and summarize this distribution using random-effects meta-analysis. We use this evidence to assess where each research area falls on the bias-variance tradeoff identified in the preceding section. These estimates can be used to help researchers make context-specific decisions about survey question formats.

We attempted a census of all list experiments ever conducted, published or otherwise, as of December 31st, 2017. We certainly failed in this task. At a minimum, we have heard from colleagues of list experiments they ran that were never written up and whose data is long since lost. We searched Google Scholar, SSRN, Dataverse, and political science conference programs for the past seven years with the search terms "list experiment," "item count technique," and "unmatched count technique," and their abbreviations. Our search yielded 487 distinct experiments in 154 separate papers. We were able to obtain both direct question and list experiment estimates in 285 cases. We limit all analyses to the 264 (92 papers) list experiments whose sensitive item was predicted by the researchers to be over- or under-reported, excluding non-sensitive traits.

We gathered statistical information about the list experiments using a cascading data collection approach. In the best case, we obtained the replication dataset from online journal appendices, Dataverse, authors' personal websites, or private communication. When replication data were not available, we searched the paper for a crosstab of list experiment responses by treatment condition (similar to Table 2). For each list experiment for which we had data or could reconstruct from a crosstab, we calculated the difference-in-means estimate of prevalence and standard error. Finally, if neither the data nor the crosstab was available, we searched the paper for the estimated prevalence rate and standard error. In rare cases, a study reported a prevalence rate estimate but no standard error we imputed our best guess based on a flexible regression model.

Publication bias arising from the file-drawer problem has, anecdotally, been a problem in the list experiment literature. In the course of our data collection, we heard from many scholars who claimed to have "not found anything" when using a list experiment. We tried hard to overcome the publication filter bias by seeking both published and unpublished experiments.

The direct question estimates of prevalence all come from the original authors. Some studies asked the direct question to either their entire sample or a random subset; others referred to a direct estimate obtained by others. We logged whichever direct estimate was reported by the original authors. We elected not to independently obtain direct prevalence estimates (e.g., from publicly-available surveys) as such discretion could lead to the perception that we were seeking to obtain a pattern either favorable or unfavorable to list experiments. We acknowledge that relying on original authors about whether and which direct questions to report introduces a second source of selection.

Meta-Analysis of Sensitivity Bias

Our measure of sensitivity bias is the difference between the list and direct estimates. We estimated the standard error of the difference as $\mathrm{SE}(\mathrm{difference}) = \sqrt{\mathrm{SE}(\mathrm{list})^2 + \mathrm{SE}(\mathrm{direct})^2}$. This formula assumes that the direct and list estimates are independent; this assumption will be mildly violated if both the direct and list estimates are calculated on the same sample. Under the assumption that direct and list estimates are positively correlated, our naive estimates of sampling variability are conservative by the properties of the variance of the difference in two random variables. We calculated a 95% confidence interval for the difference under a normal approximation.

We first categorized studies by substantive domain. This was both a practical choice (our estimator breaks down if the number of studies included is smaller than four) and a substantive choice (we think it is most useful to summarize literatures with a large amount of evidence). Second, we categorized the questions according to the expected direction of sensitivity bias: overreporting or underreporting. Wherever possible, we relied on the logics of misreporting forwarded by the original authors and in rare cases had to substitute our own best judgment. Theoretically speaking, the direction of sensitivity bias need not be constant across respondents (Lax et al. 2016), though in the vast majority of the empirical papers we reviewed, the bias was presumed to have the same sign (if not the same magnitude) for all subjects.

To summarize the distribution of estimated differences, we implement a standard random-effects meta-analysis model (DerSimonian and Laird 1986). We model observed differences y between list and direct with a normal distribution: $y \sim \mathcal{N}(\delta, \sigma)$ where σ is the observed standard error. δ represents the true sensitivity bias for a given literature and is distributed $\delta \sim \mathcal{N}(\mu, \tau)$. The moments of this distribution are μ, the grand mean level of sensitivity bias and τ, is the standard deviation of true effect sizes. We conduct Bayesian estimation using `Stan` (Carpenter et al. 2017), adopting the default improper uniform priors for θ, μ and τ, and restricting τ to be non-negative. We assess convergence by running four parallel chains and using the standard R-hat criterion. We calculate several quantities of interest from this model. First, we estimate the average amount of sensitivity bias (μ), its standard error, and 95% credible interval. Second, we estimate the *distribution* of sensitivity bias, not just its mean, since the level of bias should vary across context and topic. We calculate predictive intervals that bracket our best posterior guesses of the middle 50% and 95% of the distribution of true sensitivity biases. These ranges help us to characterize what the corpus of list experiments conducted to date teaches us about typical levels of sensitivity bias across contexts.

In order to interpret the difference between list experiments and direct questions as a measure of sensitivity bias, we make several auxiliary assumptions in addition to the standard assumptions of the list experiment. We assume no differential nonresponse between questions. We assume there are no order effects. We assume that differences in question wording of the sensitive item do not affect responses. Finally, we assume that the list experiment and direct question were asked of the same sample, or of two samples from the same population. If these additional assumptions are violated, the difference is still meaningful, but the difference itself can no longer be considered an estimate of sensitivity bias. If readers are unwilling to make these auxiliary assumptions, then our meta-analysis is still of use as a summary of how much the two measurement technologies differ.

Meta-Analysis Results

We present three sets of results. First, we summarize the estimated level of sensitivity bias in the four research literatures discussed above: vote buying, voter turnout, prejudice on the basis of race, religion, and sexual orientation, and support for authoritarian regimes. We present the study-level estimates, the meta-analysis estimate, and the predictive interval in each case. Second, we analyze all studies for which we have sufficient information according to whether authors predicted sensitivity bias in the form of overreporting or underreporting. These two analyses allow us to answer the question of whether we should worry about sensitivity bias in a given research context. Third, we

		Average Sensitivity Bias $\widehat{\mu}$		Predictive Intervals $\mathcal{N}(\widehat{\mu}, \widehat{\tau})$		
	Prediction	Estimate (s.e.)	95% C.I.	50%	95%	N studies
Vote buying	Underreporting	-0.08 (0.03)	[-0.13, -0.03]	[-0.14, -0.02]	[-0.25, 0.09]	19
Turnout	Overreporting	0.07 (0.04)	[-0.01, 0.14]	[0.00, 0.13]	[-0.12, 0.25]	10
Racial prejudice	Underreporting	0.04 (0.03)	[-0.03, 0.09]	[0.01, 0.07]	[-0.05, 0.13]	9
Religious prejudice	Underreporting	-0.01 (0.03)	[-0.08, 0.05]	[-0.06, 0.03]	[-0.14, 0.11]	12
Sexual orientation prejudice	Underreporting	0.02 (0.02)	[-0.03, 0.06]	[-0.02, 0.05]	[-0.09, 0.12]	16
	Overreporting	-0.02 (0.09)	[-0.19, 0.16]	[-0.12, 0.09]	[-0.31, 0.28]	5
Support for authoritarian regimes	Underreporting	-0.08 (0.04)	[-0.16, -0.00]	[-0.16, -0.00]	[-0.32, 0.15]	13
	Overreporting	0.14 (0.04)	[0.07, 0.21]	[0.05, 0.24]	[-0.14, 0.42]	21
All results	Underreporting	-0.03 (0.01)	[-0.05, -0.01]	[-0.12, 0.05]	[-0.27, 0.20]	196
	Overreporting	0.12 (0.02)	[0.08, 0.15]	[0.03, 0.20]	[-0.13, 0.36]	68

Table 3: Meta-analysis estimates of sensitivity bias. We include all studies for which we can estimate the sensitivity bias in the meta-analytic estimates for overreporting and for underreporting. We do not break out studies for other categories, which all have fewer than three studies.

integrate our empirical results with the design advice given above to describe where the typical study appears to fall on the bias-variance tradeoff, allowing us to answer the question of whether list experiments or direct questions are a better choice in a specific research scenario. We present our full results in Figure 3 and summarize them in Table 3.

Sensitivity Bias in Four Political Science Literatures

Clientelism in developing countries We begin our summary of results with the literature on clientelism. Across 19 studies, we find evidence of moderate underreporting of vote buying. We display this result in Figure 3 (a), which we walk through slowly for this first example.[6] The top panel of the subfigure presents the estimated level of sensitivity bias for each study (black points), calculated by subtracting the list experiment estimate from the direct question estimate. Under the assumptions laid out above, negative values indicate that the list experiment recovered a higher prevalence rate than the direct question, indicating underreporting due to sensitivity bias. For example, in the top row of Figure 3 (a), sensitivity bias is estimated to be 5 percentage points (95% CI -3 to +14) based a 2010 survey conducted in Bolivia in which the sensitive question read "They gave

[6] Full-size versions of all figures that include additional study information are included in the Supplementary Materials.

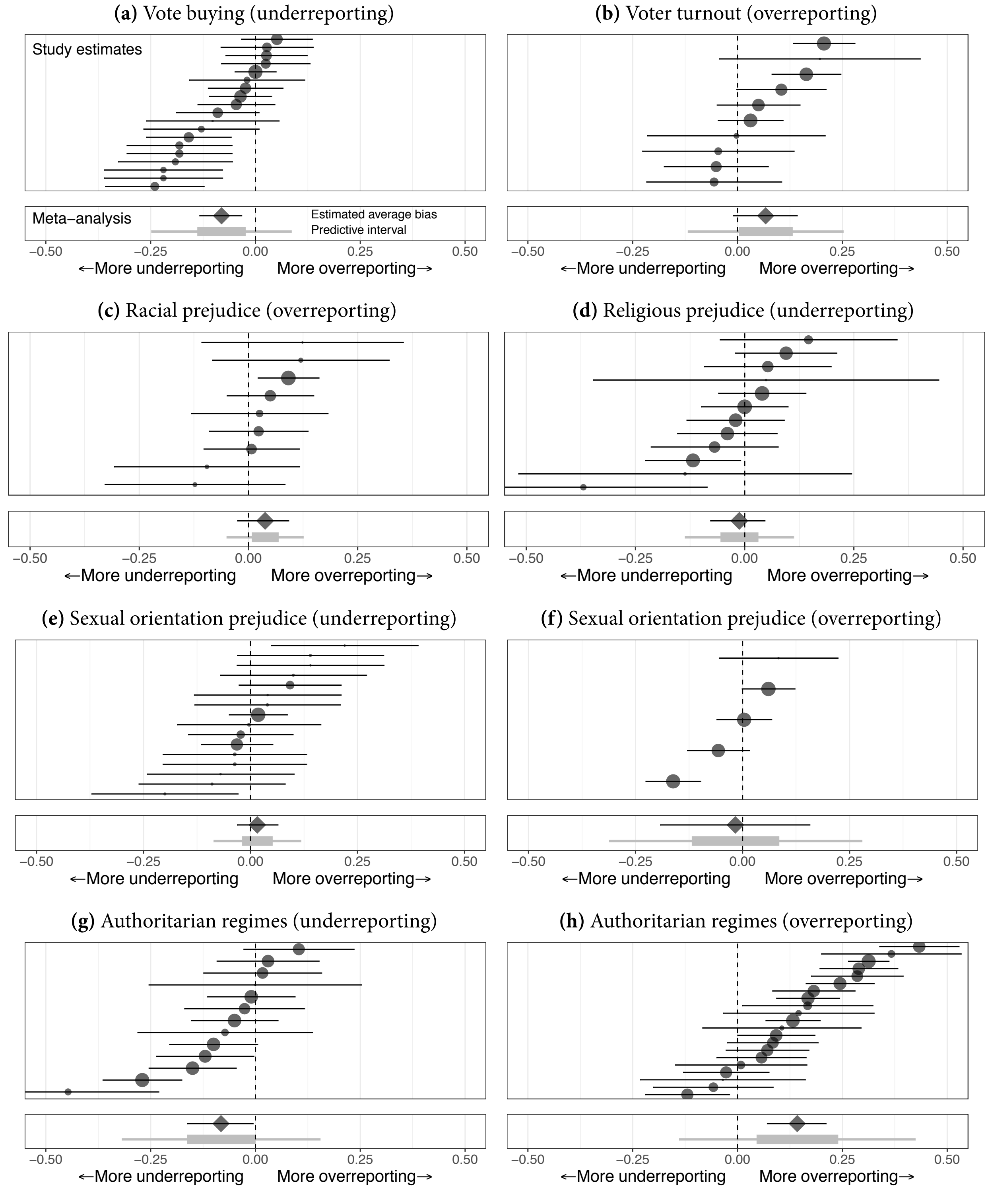


Figure 3: Sensitivity Bias in Four Political Science Research Literatures. Estimated sensitivity bias in each study with 95% confidence intervals, with point size proportional to study weight in meta-analysis (top panel). Estimated average sensitivity bias in each literature (diamond) with 95% credible interval (bottom panel) and estimated 50% (thick gray line) and 95% (thin gray line) predictive intervals for sensitivity bias.

you a gift or did you a favor" (Kiewiet de Jonge 2015). The size of the plotted points is proportional to the weight the study is accorded in the meta-analysis.[7] The size of the point represents how informative the study is, which is also reflected in the width of the confidence interval: the wider the confidence interval, the smaller the point.

We present two summaries of these estimates of sensitivity bias in the bottom panel. In the top row is the estimated average of sensitivity bias across studies for vote buying questions (black diamond), -8 points with a 95% credible interval stretching from -13 to -3 points. This is our best guess of level of sensitivity bias that would be found in a future studies absent additional contextual information. As is clear from the dispersion of the study-level estimates that range from -24 points to +5 points, the sensitivity bias attending to vote buying questions differs from context to context. In the bottom row of Figure 3 (a), we show the 50% predictive interval from -14 to -2 points and the 95% predictive interval from -27 to +10 points. Again, these intervals are different from confidence intervals in that they describe our best guess about the distribution of sensitivity biases in vote buying questions, and not our uncertainty about the *average* level of bias. In summary, the theoretical prediction of underreporting bias in direct questions about vote buying is supported on average, but there is also a considerable range of bias from very large to none at all.

Voter turnout Consistent with theory and the evidence from explicit validation studies, we do find mild evidence of overreporting voter turnout. We estimate the average bias to be +7 percentage points, though even after pooling together 10 studies, the standard error is still quite substantial, at +4 points. This uncertainty is also reflected in the very wide predictive intervals: the 50% intervals is 13 points wide and the 95% interval is 45 points wide. We interpret this evidence to indicate that at least some of the measurement error that others have documented by comparing survey responses to validated turnout records from the voter file is due to sensitivity bias, and not just memory or recall failures.

[7] The weights are calculated as, $\frac{1}{\sigma^2+\hat{\tau}^2}$, where σ^2 is the square of the observed standard error of the study and $\hat{\tau}^2$ is the estimated variance of the true sensitivity bias across studies.

Prejudice After the study of drug use Miller (1984), one of the earliest uses of the list experiment was the study of prejudice, specifically prejudice on the basis of race (Sniderman et al. 1991). Since then, list experiments have been used to study prejudice towards many subgroups within society. We measure sensitivity bias in three domains: prejudice on the basis of race, religion, and sexual orientation.[8] Contrary to expectations, we find relatively little evidence of bias, at least for the specific set of direct questions that have been tested. We were frankly quite surprised at the low levels of sensitivity bias we estimated for all three forms of prejudice.

For questions about racial prejudice, our summary guess is that if anything, subjects *overreport* racist attitudes by approximately 4 points (95% CI -3 to +9). Over the 9 studies in our sample, therefore, the difference between direct questions and list experiments is not statistically significant. The 50% predictive interval reaches from +1 to +7 points. We were surprised by this result. However, the 95% predictive interval admits large negative biases (-16 points) to large positive bias up to 16 points. Our analysis does include the 1994 Multi-Investigator study (Sniderman et al. 1994) which estimated underreporting on the scale of -10 percentage points for policy attitude questions, but +10 percentage points for non-policy attitudes like interracial dating and a black family moving in next door. Our interpretation is that either the list experiment does not provide the cover it is designed to provide in this context or that respondents who report the socially-desirable answer actually do hold the socially desirable view (at least on the narrow attitudes measured by these direct questions.) We also note that the extreme variability of the list experiment discussed above, even when tamed somewhat through meta-analysis, holds us back from drawing strong conclusions here.

Our meta-analysis again renders a null result for sensitivity bias on questions about religious prejudice. On average, we estimate a -1 point underreporting bias in direct questions thought to be prone to underreporting bias. This estimate is in the expected direction, but the credible interval is 13 points wide and includes zero. The expected range of true effects is on par with the other prejudice-

[8] The focus on these three categories instead of other, equally important subdivisions such as gender or class is entirely due to data availability. For technical reasons, our meta-analysis procedure requires a minimum of three studies in order to yield estimates of all our quantities of interest (Gelman and Hill 2006, p. 431).

related sensitivity bias estimates. Biases on the order of approximately 5 points are consistent with the set of studies in our meta-analysis.

Our set of studies includes two kinds of questions that measure attitudes towards gays and lesbians (all of the studies in our sample use questions specifically about gays and lesbians or same sex marriage and are not about the broader LGBTQ community). For questions thought to be subject to overreporting, the average estimate of sensitivity bias is -2 percentage points; for underreporting, the estimate is +2 percentage points. These estimates both have the "wrong" sign and are not distinguishable from zero. The range of plausible sensitivity biases in this literature are on the scale of 5 to 10 points.

Support for authoritarian regimes Finally, we do find evidence of substantial sensitivity bias when respondents are asked about support for authoritarian regimes and their leaders. Estimates of overreporting range up to a maximum +43 points when overreporting is predicted (Kalinin 2015) and a minimum -45 points when underreporting is predicted (Weghorst 2015). Based on 21 studies, our meta-analysis estimate of the average level for studies in which overreporting is predicted is +14 points and the 50% predictive interval ranges suggests a likely level of sensitivity bias between +4 to +24 points. When underreporting is predicted, the meta-analysis average based on 13 studies is -8 points with a 50% credible interval between -16 to 0 points. Support for authoritarian regimes is an area where, unlike others, our data suggest there is considerable risk of sensitivity bias. As we discussed in the theory section, the risks to responding to surveys in authoritarian contexts (especially on questions about politics and the regime itself) go far beyond the desire to please the interviewer. The regime is a relevant social referent and the costs range up to imprisonment or disappearance.

Sensitivity Bias by Predicted Direction of Misreporting

In the previous section, we zoomed in on results from four political science literatures. In this section, we zoom out to the full set of studies for which we have both list and direct estimates, regardless of discipline or topic. Figure 4 plots the direct question estimate against the list experiment estimate separately by the predicted direction of sensitivity bias. The point size is proportional to the standard error of the difference estimate (more precise estimates are larger). We also present 95% confidence intervals for both estimates. The regression line overlaid on top of the raw estimates is fit via Deming regression (Deming 1943), an errors-in-variables model, which is appropriate given the measurement error in both the left-hand and right-hand sides of the equation. We estimate measurement error with the standard errors of the direct and list estimates.

First, we see that the direct and list estimates are highly correlated – prima facie evidence that whatever the measurement properties of direct questions and list experimentation, they appear to be measure the same latent quantity. One measure of the strength of this correlation is the slope of the Deming regressions, both of which are close to 1. Second, as shown in Table 3, the average sensitivity bias in the case of underreporting is -3 points (SE: 1 point). For overreporting, the bias is much larger at +12 points (SE: 2 points). This asymmetry can be observed by comparing the two facets of Figure 4. For overreporting, points lie overwhelmingly above the 45 degree line, whereas for underreporting points cluster tightly around it.

Empirical distribution of sensitivity bias and sample size

Our final set of results uses the empirical distribution of sensitivity bias as a means to gauge the extent to which list experiments conducted to date are of a sufficiently large size. We return to the two main goals of list experimentation: achieving a better RMSE or demonstrating the existence of sensitivity bias.

The graph shows three regions. Below the lower curve (73 of 187 studies), it is likely that direct questioning would have produced answers closer to the truth (in RMSE terms) than the list experiments. Between the two curves (78 studies), the choice between list experiments and the direct question depends on the goal of the research. These list experiments are large enough to produce lower RMSE than the direct question, but are not large enough to reliably demonstrate the existence of sensitivity bias. The studies that are above both curves (36 studies) are large enough to be

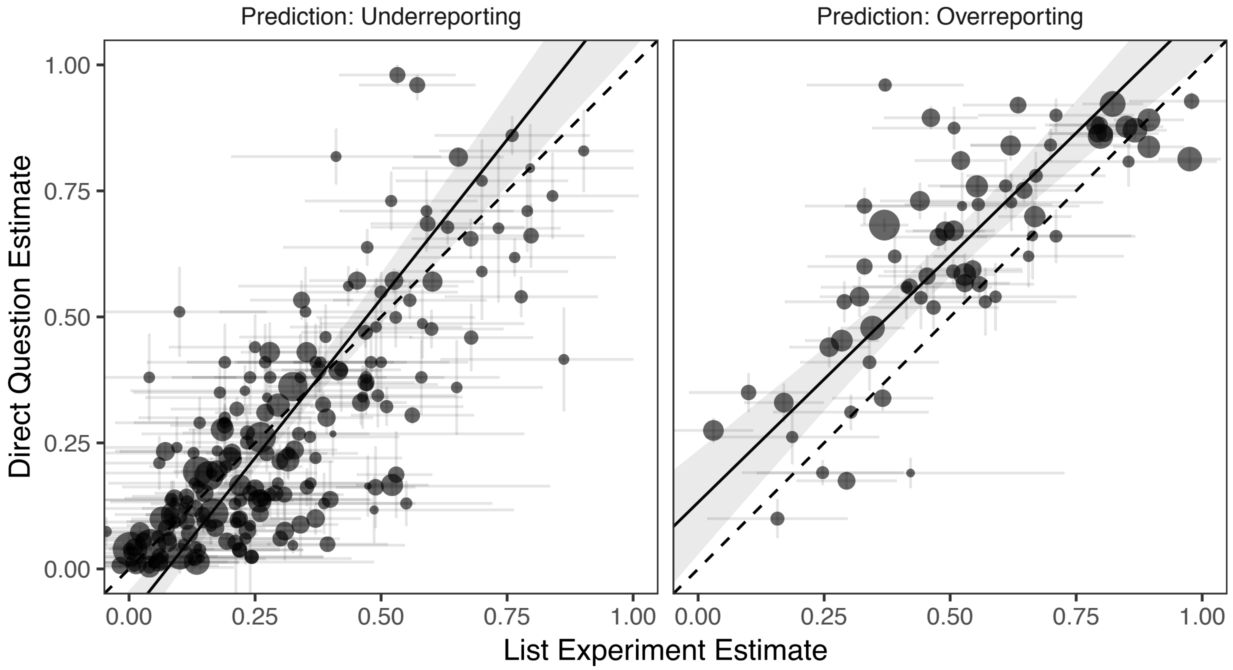


Figure 4: List Experiment Estimates are Correlated with Direct Question Estimates, and Across Domains There is Sensitivity Bias Especially when Overreporting is Predicted. Estimates of the prevalence rate of the sensitive item from the list experiment (x axis) and from a direct question (y axis) are presented as points along with 95% confidence intervals of each estimate (light gray lines) with point size proportional to the weight from a Deming errors-in-variables regression. The Deming regression model fit (solid line) is presented along with its 95% confidence interval (gray area). The 45% degree line, representing no sensitivity bias is plotted as a dashed line.

preferable for either purpose. Figure 5 shows that many list experiments are simply too small.

We emphasize that the indifference curves between list experiments and direct questions included in Figure 5 assume the standard list experiment design. The true position of each study relative to indifference between the two designs is better represented by its effective sample size, adjusting for any improvements to list experiment design and analysis implemented in that study. In the supplementary materials, we present effective sample sizes for a set of possible design and analysis modifications.

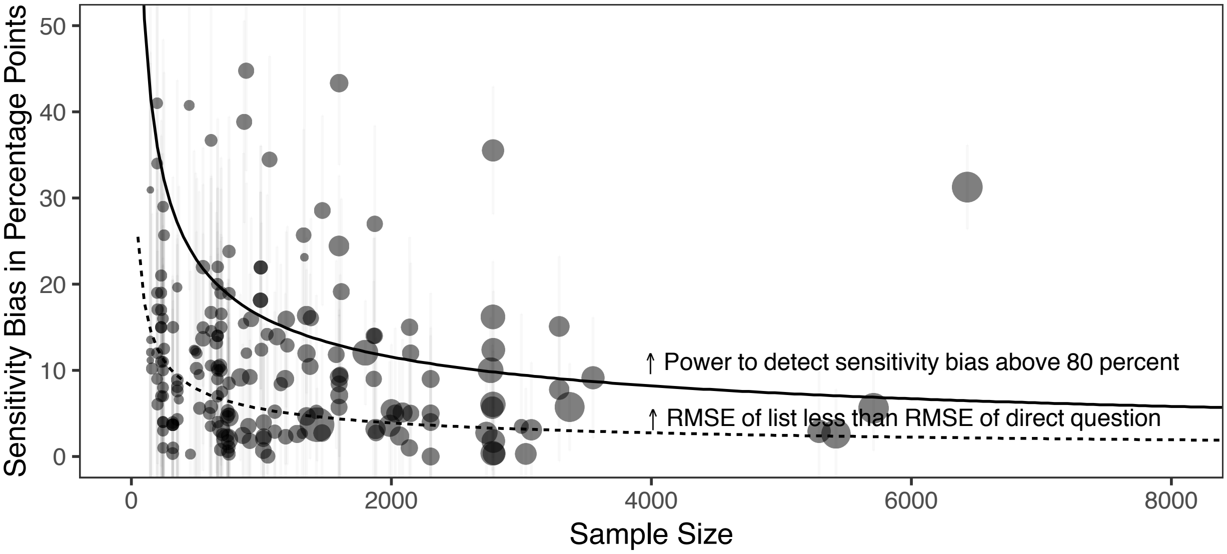


Figure 5: Many Studies of Sensitive Topics Are Smaller than Sample Sizes Recommended based on Power or Root Mean-Squared Error Tradeoffs. Existing studies are overlaid on two design recommendations: the power indifference curve for designs estimating the amount of social desirability bias, and the RMSE indifference curve for designs estimating the prevalence rate.

Summary of Empirical Results

Is sensitivity bias likely to be a problem? Perhaps unsurprisingly given the huge range of questions that have been investigated using list experiments over the past three decades, the answer is, "it depends." Subjects substantially overreport support for authoritarian regimes, underreport opposition to them, and underreport vote buying. We find moderate evidence of overreporting of voter turnout. Surprisingly to us, subjects appear to report honestly answers to questions about prejudice on the basis of race, religion, and sexual orientation.

Our meta-analysis faces some important limitations. First and foremost, this is not a validation study since for most topics, we do not have access to the true prevalence rate. Indeed, this lack is what occasions the reliance of survey estimates of prevalence in the first place. As a result, the results can be interpreted differently, depending on the assumptions one is willing to make. If the list experiment assumptions (no liars and no design effects) hold, we can interpret the estimated differences between direct and list responses as an estimate of sensitivity bias. If these assumptions do not hold, the difference between list and direct estimates simply represents the difference in the answer obtained depending on which measurement technology is used. The difference may represent the change to respondent beliefs about which social referents can obtain answers between the list experiment and the direct question. This quantity is still important to learn: given the popularity of the list experiment and other related measurement techniques, many researchers are faced with the choice of asking directly or using another method. Our estimates can help guide those choices even if the meta-estimates do not represent unbiased estimates sensitivity bias per se.

Another limitation concerns the variability of the list experiment. The power of the list experiment to detect moderate sensitivity bias is low, so our conclusion of limited bias in most direct measures may be an instance of "accepting the null" of no bias. The more cautious interpretation is that we can rule out average biases greater than 10 or 15 percentage points in most cases. Biases on this order are of course very meaningful, but also difficult to detect with list experiments. The posterior predictive intervals are wider, which indicate that the biases in some contexts could be much larger than the average bias, which may justify the use of list experiments.

Despite the reasonable concern that the list experiment assumptions are unlikely to hold in at least some contexts, the technology appears to perform surprisingly well. In the 166 list experiments for which we have sufficient information, 160 pass the design effects test. The tool itself appears to generate approximately unbiased answers. However, we highlight an important limitation of our meta-analysis search: we are likely to miss list experiments for which the design assumptions are violated due to publication bias. If there is a large corpus of these failed list experiments, our conclusions might differ.

At the same time, list experiments are known to be high variance; our analysis reveals that most list experiments that have been conducted thus far have probably been too small for their research goals. Increasing sample sizes can of course be very expensive and researchers must weigh the data costs against the benefits of generating precise answers to their research questions, but the field should clearly move away from conducting underpowered list experiments. If the expected bias is below 10 percentage points, direct questions (if ethically and operationally feasible) should be preferred if sample sizes are less than 3,000. Where possible, the design innovations described above should be incorporated.

From a research design perspective, the moderate levels of sensitivity bias our meta-analysis identifies presents scholars with a difficult choice. Each technology for addressing sensitivity bias comes with costs, and that cost is often a loss of statistical power. The high variance of list experiments means that the substantive conclusions will be highly uncertain. In terms of mean squared error, *unbiased* list experiments are dominated by *biased* direct questions at most sample sizes. In other words, unless sample sizes exceed 3,000 subjects, direct questions will tend to be more accurate in terms of RMSE than list experiments, even in the presence of non-negligible sensitivity biases. In settings where alternative techniques like the list experiment are appropriate, researchers should combat the high costs of these techniques with state-of-the-art designs and increased sample sizes far beyond those used in current practice.

Discussion

Survey research designs rely on asking respondents for self-reports of political attitudes, beliefs, and behaviors. When respondents refuse to respond, or answer but misreport, the conclusions from survey data will be biased. In this paper, we set out to answer two questions: How much of a problem is sensitivity bias and what can be done about it.

With respect to the first question, we think researchers should ask themselves four questions when deciding whether to worry about the problem. (1) Is there a social referent respondents have in mind when answering? (2) Do respondents believe the social referent can obtain their answers? (3) Do respondents perceive that the social referent prefers a particular answer to the question (4) Do respondents believe they (or others) will suffer costs if that preferred response is not provided? If the answer to any of these questions is “no,” then sensitivity bias may not be a meaningful source of measurement error. Researchers may be uncertain as to the answer to each of these questions, in which case care and caution are of course still warranted.

With respect to what researchers should do about sensitivity bias, we characterized the choice between list experiments and direct questions as a bias-variance tradeoff and calculated the conditions under which one technology would be preferred to the other. Under typical conditions, list experiments are approximately seven times noisier than direct questions, which means that either the sample size or the amount of bias needs to be large in order to justify a list experiment.

Beyond the list experiment, many techniques have been proposed to mitigate sensitivity bias, and the social reference theory helps us to consider these alternative measurement procedures in a common framework. They largely fall into two types. The first set combine responses with random noise in order to change respondent beliefs about which social referents can obtain their responses. This set includes the list experiment, the randomized response technique, and the crosswise technique. These techniques obscure responses both from interviewers and bystanders at the time of the survey, but also from later data consumers including researchers and state authorities. If these are the social referents respondents are worried demand a particular response and will exact costs if it is not provided, these methods will reduce sensitivity bias.

The second set of techniques changes the context in which the sensitive question is asked to separate the response from the identity of the respondent temporarily. Most simply, researchers have implemented secret ballots during surveys to measure vote choice or turnout (Bishop and Fisher 1995), physical barriers between respondents and interviewers (Scacco 2012), and questionnaires recorded on MP3 players with item order randomized (Chauchard 2013). Self-administered surveys and interactive voice response surveys (i.e., with no interviewer involved at all) are often deployed for sensitive surveys for similar reasons. These procedures only obscure responses temporarily, preventing bystanders and in some cases interviewers from linking responses to particular responses. The researchers and state authorities who intercept the data can reconnect responses to identifying information.[9] When interviewers or bystanders are the primary social referents respondents worry about, these methods will reduce sensitivity bias as well.

Our research design advice can also guide decision-making between these techniques. All of the techniques that reduce sensitivity bias by adding noise will exhibit a bias-variance tradeoff similar to the one we described for the list experiment. By design, the other information (random noise or unrelated items) reduces precision compared to the direct question. The techniques that temporarily separate respondents and responses do not typically face a bias-variance tradeoff – they decrease bias without incurring variance penalties. When these techniques can be ethically and logistically deployed (and when the interviewer or bystanders are the primary concern of respondents) they may be the best choice to address sensitivity bias.

We have focused on latent attitudes, beliefs, and behaviors that respondents know (or could know) they hold or have taken. In other words, our target of inference is the binary latent trait that we formalized as D_i^*. When designed with appropriate construct validity, direct questions, list experiments, and the randomized response technique all target this quantity, with varying degrees of bias depending on the extent of sensitivity bias. Other techniques, such as the implicit association test (Greenwald et al. 1998) and the affect misattribution procedure (Payne et al. 2005), target implicit attitudes. Because it is currently unclear whether (or the extent to which) responses to implicit attitudes measurement techniques are under the conscious control of subjects (Nosek 2007), it is also unclear whether such measures are plausibly distorted by sensitivity bias (though see Skorinko and Sinclair 2018).

In this paper, we have shown that sensitivity bias is a very real concern for survey research, but it varies considerably by context and by topic in sometimes surprising ways. In some literatures, we find little to no evidence of sensitivity bias and in others it is quite sizable. We hope researchers will carefully consider the likely magnitude of sensitivity biases using the four sensitivity criteria

[9] This need not be the case. For an exception, see Scacco (2012), in which sensitive responses were linked to individuals only through a randomized code that could only be reconnected at the researcher's home institution. This was designed to prevent state authorities from connecting responses about participation in riots.

before turning to the list experiment or similar techniques, mainly because they are so variable that achieving good precision can be costly. Surveys may also be substantially improved by less expensive changes to survey administration that can reduce bias without increasing variance. When list experiments or similar methods are selected, they should be conducted only with large samples or when biases are expected to be substantial.

References

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437–459.

Aquilino, William S. 1993. "Effects of spouse presence during the interview on survey responses concerning marriage." *Public Opinion Quarterly* 57(3):358–376.

Belli, Robert F., Michael W. Traugott, Margaret Young and Katherine A. McGonagle. 1999. "Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring." *The Public Opinion Quarterly* 63(1):90–108.

Berinsky, Adam J. 2004. "Can we talk? Self-presentation and the survey response." *Political Psychology* 25(4):643–659.

Bishop, George F. and Bonnie S. Fisher. 1995. ""Secret ballots" and self-reports in an exit-poll experiment." *Public Opinion Quarterly* 59(4):568–588.

Blaydes, Lisa and Rachel M. Gillum. 2013. "Religiosity-of-Interviewer Effects: Assessing the Impact of Veiled Enumerators on Survey Response in Egypt." *Politics and Religion* 6(3):459–482.

Bobo, Lawrence. 1988. Group conflict, prejudice, and the paradox of contemporary racial attitudes. In *Eliminating racism*. Springer pp. 85–114.

Brusco, Valeria, Marcelo Nazareno and Susan Carol Stokes. 2004. "Vote buying in Argentina." *Latin American Research Review* 39(2):66–88.

Burden, Barry C. 2000. "Voter turnout and the national election studies." *Political Analysis* 8(4):389–398.

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

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan: A probabilistic programming language." *Journal of statistical software* 76(1).

Catania, Joseph A., Diane Binson, Jesse Canchola, Lance M. Pollack, Walter Hauck and Thomas J. Coates. 1996. "Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior." *Public Opinion Quarterly* 60(3):345–375.

Chauchard, Simon. 2013. "Using MP3 players in surveys: The impact of a low-tech self-administration mode on reporting of sensitive attitudes." *Public Opinion Quarterly* 77(S1):220–231.

Chen, Yuyu and David Y. Yang. 2018. "The Impact of Media Censorship: Evidence from a Field Experiment in China." Working paper, Applied Economics Department, Guanghua School of Management.

Corstange, Daniel. 2014. "Foreign-Sponsorship Effects in Developing-World Surveys: Evidence from a Field Experiment in Lebanon." *Public Opinion Quarterly* 78(2):474–484.

Corstange, Daniel. 2017. "Clientelism in Competitive and Uncompetitive Elections." *Comparative Political Studies* 51(1):76–104.

Cotter, Patrick R., Jeffrey Cohen and Philip B. Coulter. 1982. "Race-of-interviewer effects in telephone interviews." *Public Opinion Quarterly* 46(2):278–284.

Cruz, Cesi. 2019. "Social Networks and the Targeting of Vote Buying." *Comparative Political Studies* 52(3):382–411.

Davis, Darren W. 1997. "The direction of race of interviewer effects among African-Americans: Donning the black mask." *American Journal of Political Science* pp. 309–322.

Deming, William Edwards. 1943. *Statistical Adjustment of Data*. Wiley.

DerSimonian, Rebecca and Nan Laird. 1986. "Meta-analysis in clinical trials." *Controlled clinical trials* 7(3):177–188.

Enamorado, Ted and Kosuke Imai. 2018. "Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records.".

Feldman, Jacob J., Herbert Hyman and Clyde W. Hart. 1951. "A field study of interviewer effects on the quality of survey data." *Public Opinion Quarterly* 15(4):734–761.

Fisher, Robert J. 1993. "Social desirability bias and the validity of indirect questioning." *Journal of consumer research* 20(2):303–315.

Frye, Timothy, Ora John Reuter and David Szakonyi. 2014. "Political Machines at Work Voter Mobilization and electoral subversion in the Workplace." *World Politics* 66(2):195–228.

Frye, Timothy, Scott Gehlbach, Kyle L Marquardt and Ora John Reuter. 2017. "Is Putin's popularity real?" *Post-Soviet Affairs* 33(1):1–15.

Gandhi, Jennifer and Ellen Lust-Okar. 2009. "Elections under authoritarianism." *Annual review of political science* 12:403–422.

Gelman, Andrew and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gervais, Will M. and Maxine B. Najle. 2018. "How many atheists are there?" *Social Psychological and Personality Science* 9(1):3–10.

Gilens, Martin. 2009. *Why Americans hate welfare: Race, media, and the politics of antipoverty policy*. University of Chicago Press.

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.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1):159–172.

Goffman, Erving. 1959. *The presentation of self in everyday life*. Anchor Books.

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.

González-Ocantos, Ezequiel, Chad Kiewiet de Jonge and David W Nickerson. 2015. "Legitimacy buying: The dynamics of clientelism in the face of legitimacy challenges." *Comparative Political Studies* 48(9):1127–1158.

Greenwald, Anthony G., Debbie E. McGhee and Jordan L.K. Schwartz. 1998. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of Personality and Social Psychology* 74(6):1464.

Greenwald, Anthony G. and Mahzarin R. Banaji. 1995. "Implicit social cognition: attitudes, self-esteem, and stereotypes." *Psychological Review* 102(1):4.

Greenwald, Anthony G. and Steven J. Breckler. 1985. To whom is the self presented. In *The Self and Social Life*, ed. Barry R. Schlenker. New York: McGraw-Hill pp. 126–145.

Haire, Mason. 1950. "Projective techniques in marketing research." *Journal of Marketing* 14(5):649–656.

Hatchett, Shirley and Howard Schuman. 1975. "White respondents and race-of-interviewer effects." *The Public Opinion Quarterly* 39(4):523–528.

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.

Huddy, Leonie, Joshua Billig, John Bracciodieta, Lois Hoeffler, Patrick J. Moynihan and Patricia Pugliani. 1997. "The effect of interviewer gender on the survey response." *Political Behavior* 19(3):197–220.

Imai, Kosuke, Bethany Park and Kenneth F. Greene. 2014. "Using the predicted responses from list experiments as explanatory variables in regression models." *Political Analysis* 23(2):180–196.

Kalinin, Kirill. 2015. "Exploring Putin's Post-Crimean Supermajority.".

Kane, Emily W. and Laura J. Macaulay. 1993. "Interviewer gender and gender attitudes." *Public opinion quarterly* 57(1):1–28.

Kiewiet de Jonge, Chad P. 2015. "Who lies about electoral gifts? Experimental evidence from Latin America." *Public Opinion Quarterly* 79(3):710–739.

Kinder, Donald R and David O. Sears. 1981. "Prejudice and politics: Symbolic racism versus racial threats to the good life." *Journal of personality and social psychology* 40(3):414.

Kramon, Eric. 2016. "Where is vote buying effective? Evidence from a list experiment in Kenya." *Electoral Studies* 44:397–408.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European revolution of 1989." *World politics* 44(1):7–48.

Kuran, Timur. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

Lax, Jeffrey R., Justin Phillips and Alissa F. Stollwerk. 2016. "Are Survey Respondents Lying About their Support for Same-Sex Marriage? Lessons from A Recent List Experiment." *Public Opinion Quarterly* 80(2):510—533.

Leary, Mark R. and Robin M. Kowalski. 1990. "Impression management: A literature review and two-component model." *Psychological bulletin* 107(1):34.

Littman, Rebecca. 2015. "A Challenge for Psychologists: How to Collect Sensitive Information in Field Experiments." International Society of Political Psychology Blog.

Lohmann, Susanne. 1994. "The dynamics of informational cascades: the Monday demonstrations in Leipzig, East Germany, 1989–91." *World politics* 47(1):42–101.

Maccoby, Eleanor E. and Nathan Maccoby. 1954. "The interview: A tool of social science." *Handbook of social psychology* 1:449–487.

Mares, Isabela, Aurelian Muntean and Tsveta Petrova. 2016. "Economic Intimidation in Contemporary Elections: Evidence from Romania and Bulgaria." *Government and Opposition* pp. 1–32.

Mares, Isabela and Lauren E Young. 2016. The core voter's curse: Coercion and clientelism in Hungarian elections. Technical report Working paper, Columbia University.

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

Nosek, Brian A. 2007. "Implicit–explicit relations." *Current Directions in Psychological Science* 16(2):65–69.

Ostwald, Kai and Guillem Riambau. 2017. "Voting Behavior under Doubts of Ballot Secrecy.".

Paulhus, Delroy L. 1991. Measurement and control of response bias. In *Measures of social psychological attitudes*, ed. John P. Robinson, Phillip R. Shaver and Lawrence S. Wrightsman. Vol. 1 San Diego, C.A.: Academic Press pp. 17–59.

Payne, B. Keith, Clara Michelle Cheng, Olesya Govorun and Brandon D Stewart. 2005. "An inkblot for attitudes: affect misattribution as implicit measurement." *Journal of personality and social psychology* 89(3):277.

Scacco, Alexandra. 2012. "Anatomy of a Riot: Participation in Ethnic Violence in Nigeria." *Book Manuscript, New York University* .

Schuman, Howard, Charlotte Steeh, Lawrence Bobo and Maria Krysan. 1997. *Racial attitudes in America: Trends and interpretations*. Harvard University Press.

Silver, Brian D., Paul R. Abramson and Barbara A. Anderson. 1986. "The presence of others and overreporting of voting in American national elections." *Public Opinion Quarterly* 50(2):228–239.

Simpser, Alberto. 2017. "Why Do Sensitive Survey Questions Elicit Truthful Answers? Theory and Evidence with Application to the RRT and the List Experiment." Working paper.

Skorinko, Jeanine L. M. and Stacey Sinclair. 2018. "Shared reality through social tuning of implicit prejudice." *Current Opinion in Psychology* 23:109 – 112.

Sniderman, Paul M., Henry E. Brady and Philip E. Tetlock. 1994. "1994 Multi-Investigator Study." Dataset.

Sniderman, Paul M and Philip E. Tetlock. 1986. "Reflections on American racism." *Journal of Social Issues* 42(2):173–187.

Sniderman, Paul M., Philip E. Tetlock and Thomas Piazza. 1991. "National Race and Ethnic Politics Survey." Dataset.

Snyder, Mark. 1987. *Public appearances, Private realities: The psychology of self-monitoring.* W.H. Freeman.

Stokes, Susan C. 2005. "Perverse accountability: A formal model of machine politics with evidence from Argentina." *American political science review* 99(3):315–325.

Stokes, Susan C. 2007. Political clientelism. In *The Oxford handbook of political science*, ed. Carles Boix and Susan C. Stokes. Oxford: Oxford University Press.

Tang, Wenfang. 2016. *Populist authoritarianism: Chinese political culture and regime sustainability*. Oxford University Press.

Tesler, Michael. 2012. "The Return of Old-Fashioned Racism to White Americans' Partisan Preferences in the Early Obama Era." *The Journal of Politics* 75(1):110–123.

Tourangeau, Roger, Lance J. Rips and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive questions in surveys." *Psychological bulletin* 133(5):859.

Vavreck, Lynn. 2007. "The exaggerated effects of advertising on turnout: The dangers of self-reports." *Quarterly Journal of Political Science* 2(4):325–344.

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

Weghorst, Keith R. 2015. "Political Attitudes and Response Bias in Semi-Democratic Regimes.".

Weitz-Shapiro, Rebecca. 2012. "What wins votes: Why some politicians opt out of clientelism." *American Journal of Political Science* 56(3):568–583.

Wintrobe, Ronald. 2000. *The political economy of dictatorship*. Cambridge: Cambridge University Press.

Yuri Levada Analytical Center. 2019. "Putin's Approval Rating." Indicators report, 1999 to 2019.

Supplementary Materials for:

When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments

Graeme Blair[†] Alexander Coppock[‡] Margaret Moor[§]

Contents

[†] Graeme Blair is Assistant Professor of Political Science, University of California, Los Angeles. `https://graemeblair.com`

[‡] Alexander Coppock is Assistant Professor of Political Science, Yale University. `https://alexandercoppock.com`

[§] Margaret Moor is a graduate of Yale College, Class of 2018.

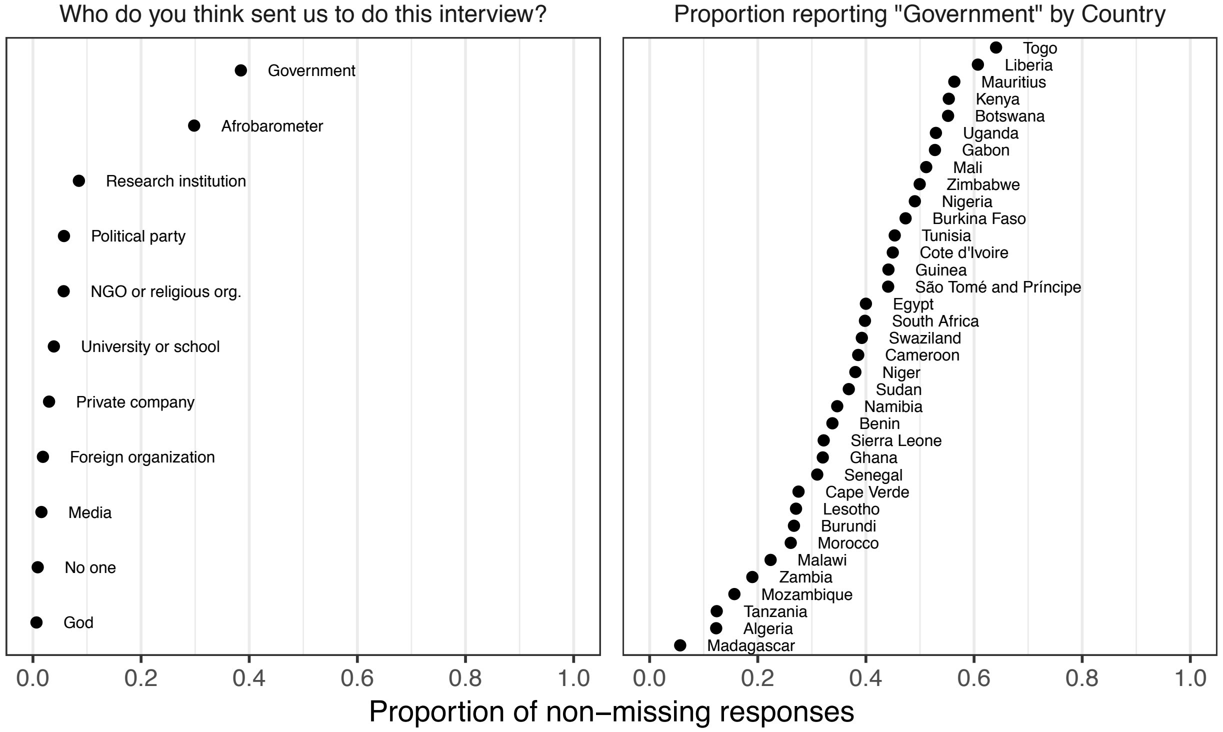


Figure 1: Beliefs about the organization who sent the interviewer vary both across individuals (left panel) and across countries (right panel). Data from Afrobarometer Round 6 (2014–2015). The question text is "Just one more question: Who do you think sent us to do this interview?" and responses were coded by Afrobarometer from recorded verbatim responses.

1. Perceptions of Survey Sponsors Vary by Individual and Context

In the Afrobarometer face-to-face survey conducted in several countries in Sub-Saharan Africa, the last question in every survey (since the second Afrobarometer round) asks who respondents think are responsible for the survey. The question text is, "`Just one more question: Who do you think sent us to do this interview?`" Responses are coded by Afrobarometer from recorded verbatim responses. In Figure 1, we display the proportion of responses to each answer option overall (left panel) and the proportion responding that the "government" is responsible for the survey across countries (right panel). The figure shows that responses vary substantially across respondents, and across countries. Impression management concerns and the perceived risks of disclosure are likely heterogeneous across respondents and contexts.

2. How the List Experiment Addresses Sensitivity Bias

The list experiment obscures individual responses to the sensitive item, but still allows analysts to estimate sample quantities including sensitive item prevalence and other relevant quantities. With a set of N individuals indexed by i, we randomly assign each to a treatment group ($T_i = 1$) or a control group ($T_i = 0$). In the control group, we ask respondents for a count of the number of "yes" responses to J control items indexed by j. In the treatment group, we ask respondents for a count of the number of "yes" responses to a set of $J + 1$ items, the J control items plus the sensitive item. We define two sets of potential outcomes: $Z_{ij}(t)$ for $t = 0, 1$. The observed outcome is defined as $Y_i = Y_i(T_i)$.

A main aim of researchers is to estimate the sample prevalence of the sensitive item: $\pi^* = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$. In order to identify this quantity, four assumptions must be invoked. These are described in Imai (2011), but we recapitulate them here. First, we need the standard assumptions for identifying the average treatment effect in an experiment: noninterference and the ignorability of the treatment status. Noninterference requires that subjects' outcomes depend only on their own treatment status and not on that of other subjects. In list experiments, noninterference is typically assured by design because subjects take the surveys separately. Ignorability requires that the treatment be independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$ and is guaranteed by design in list experiments because the treatment is randomized.

Two additional assumptions are required in order to interpret this treatment effect as the prevalence rate of the sensitive item. No design effects assumes that responses to the control items do not differ in treatment and control. This assumption would be violated if the presence of the sensitive item changes how subjects respond to the control items. Formally, the no design effects assumption states that for all respondents i, $\sum_{j=1}^{J} Z_{ij}(0) = \sum_{j=1}^{J} Z_{ij}(1)$. No liars assumes that respondents do not misreport the "yes" or "no" response to the sensitive item. The no liars assumption states that for all respondents i, $Z_{i,J+1}(1) = D_i^*$. Substantively, no liars means that list experiment responses are not distorted by sensitivity bias. The protection provided by the list experiment removes the threat of costs because the social referent cannot learn subjects' responses.

No liars might be violated if treatment group subjects' true response to the list experiment would be "all" or "none," but they report a different value instead. An answer of "none" would identify them as a "no" to the sensitive item and an answer of "all" would identify them as a "yes" to the sensitive item. For these respondents, the list experiment offers no protection from the aggregation with the control items, so we should not expect a change in the self-presentation pressures or the risk of disclosure. Glynn (2013) describes this specific violation of no liars as floor and ceiling effects. No liars would also be violated if subjects were unable to admit the truth to themselves.

Violations of no design effects occur when respondents evaluate the control items differently in treatment and control. Respondents may be affected simply by the number of items in a list, so in the treatment group which has one more item than control respondents may change responses to the control items (Flavin and Keane 2009). If respondents evaluate items in a list relative to each other, the addition of a new item may change their evaluations of the control items. Indeed, even if respondents do not evaluate items relative to one another, the addition of the sensitive item may simply act as a frame that changes how they think about other items.

Design effects may also be induced by the presence of the sensitive item in the treatment group list due to its sensitivity. Scholars worry that adding the sensitive item triggers impression management concerns generally, and that may affect responses to the control items. Zigerell (2011) notes that respondents may want to send a strong signal that they do are not answering the sensitive item in the affirmative by deflating their responses to the control items to be closer to or at a zero response.

Under noninterference, ignorability, no design effects, and no liars, the sample sensitive item prevalence is nonparametrically identified. We estimate this quantity using the difference-in-means estimator, which is an unbiased estimator under these assumptions.[1]

Other quantities beyond the sensitive item prevalence have been of interest to political scientists. Subgroup prevalence (analogous to conditional average treatment effects in standard experimental settings) and their differences can be estimated with the same tools and justifications. For surveys that also include a direct question on the same topic (such as the Kenya postelection survey reported in Kramon 2016), the difference between the list experiment and the direct question is estimate of sensitivity bias (Janus 2010; Blair and Imai 2012).

[1] The difference-in-means estimator is not the only way to estimate the prevalence rate. Other estimators, such as the nonlinear least squares and maximum likelihood procedures whose main purpose is the estimation of multiple regression coefficients, may generate more precise estimates of the prevalence rate, but do so at the cost of additional modeling assumptions (Imai 2011; Blair et al. Forthcoming).

3. Variance Derivations

Variance of the direct question estimator

In the main text, we use the following expression to describe the variance of the direct question estimator of in terms of the sample size n, the true prevalence rate π^*, and the level of sensitivity bias δ:

$$\mathbb{V}(\widehat{\pi}) = \frac{\pi^*(1-\pi^*) + \delta(1-\delta) + 2(\delta - \pi^*\delta)}{n-1}$$

Subject i's true latent trait is D_i^*. The response that subject is would give to the direct question is D_i. We define the difference between these as $W_i \equiv D_i^* - D_i$. Sensitivity bias, therefore is the expectation of W_i: $\delta = \mathbb{E}[W_i]$ The direct question estimator $\widehat{\pi}$ is the sample mean $\widehat{\pi} = \frac{1}{n}\sum_1^n D_i$, which has variance $\frac{\mathbb{V}(D_i)}{n-1}$ by standard formulas. Since $D_i = D_i^* - W_i$, the variance of D_i can be written $\mathbb{V}(D_i^*) + \mathbb{V}(W_i) + 2Cov(D_i^*, W_i)$. We need an expression for $Cov(D_i^*, W_i)$. Here we add an additional assumption of monotonicity that states that the value of W_i is either 0 or 1 for all subjects, as in the typical underreporting case. An analogous expression holdes in the overreporting case. Monotonicity may not hold in the entire subject pool, but will for well constructed subsets.

$$\begin{aligned}
Cov(D_i^*, W_i) &= \mathbb{E}[(D_i^* - \mathbb{E}[D_i^*])(W_i - \mathbb{E}[W_i])] \\
&= \mathbb{E}[(D_i^* - \pi^*)(W_i - \delta)] \\
&= \mathbb{E}[(D_i^* W_i)] - \mathbb{E}[D_i^*\delta] - \mathbb{E}[\pi^* W_i] + \mathbb{E}[\pi^*\delta] \\
&= \delta - \pi^*\delta - \pi^*\delta + \pi^*\delta \\
&= \delta - \pi^*\delta
\end{aligned}$$

Plugging this expression back in, we see that

$$\begin{aligned}
\mathbb{V}(\widehat{\pi}) &= \frac{\mathbb{V}(D_i^*) + \mathbb{V}(W_i) + 2Cov(D_i^*, W_i)}{n-1} \\
&= \frac{\pi^*(1-\pi^*) + \delta(1-\delta) + 2(\delta - \pi^*\delta)}{n-1}
\end{aligned}$$

Variance of the list experiment estimator

In the main text, we use the following expression to describe the variance of the list experiment estimator ($\widehat{\pi^*}$) of in terms of the sample size n, the true prevalence rate π^*, and the variance of the control item response $\mathbb{V}(Y_i(0))$.

$$\mathbb{V}(\widehat{\pi^*}) = \frac{4\mathbb{V}(Y_i(0) + \pi^*(1-\pi^*)}{n-1}.$$

The square of Eq. 3.4 in Gerber and Green (2012) defines the variance of the difference-in-means estimator under complete random assignment as follows.

$$\mathbb{V}(\widehat{\pi^*}) = \frac{1}{n-1}\left\{\frac{m\mathbb{V}(Y_i(0))}{n-m} + \frac{(N-m)\mathbb{V}(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right\}$$

We can rewrite this expression to suit our purposes by assuming half the units are treated ($m = n/2$) and by assuming no liars and no design effects, so $Y_i(1) = Y_i(0) + D_i^*$.

$$\begin{aligned}
\mathbb{V}(\widehat{\pi^*}) &= \frac{1}{n-1}\left\{\mathbb{V}(Y_i(0)) + \mathbb{V}(Y_i(1)) + 2Cov(Y_i(0), Y_i(1))\right\} \\
&= \frac{1}{n-1}\left\{\mathbb{V}(Y_i(0)) + \mathbb{V}(Y_i(0) + D_i^*) + 2Cov(Y_i(0), Y_i(0) + D_i^*)\right\}
\end{aligned}$$

Assume $Y_i(0) \perp D_i^*$ (and recall that if $Y \perp Z$, then $cov(X, Y+Z) = cov(X,Y) + cov(X,Z)$)). This assumption can be bolstered by design if researchers choose control items whose sum is orthogonal to the sensitive trait. If researchers follow the design advice in Glynn (2013) any covariance between the control items and the sensitive trait will be minimized.

$$\begin{aligned}
\mathbb{V}(\widehat{\pi^*}) &= \frac{1}{n-1}\left\{\mathbb{V}(Y_i(0)) + \mathbb{V}(Y_i(0)) + \mathbb{V}(D_i^*) + 2 * \mathbb{V}(Y_i(0))\right\} \\
&= \frac{4\mathbb{V}(Y_i(0)) + \mathbb{V}(D_i^*)}{n-1} \\
&= \frac{4\mathbb{V}(Y_i(0) + \pi^*(1-\pi^*)}{n-1}
\end{aligned}$$

4. Additional Simulation Evidence

Estimate Group Differences in Prevalence Rates

Many social scientific theories predict that prevalence rates will differ according to subgroups defined by individual-level covariates such as race, gender, or political orientation. Further, some experimental interventions are designed to *change* whether or not a person holds an attitude or engages in a behavior. In both the observational and experimental cases, a common concern is that the differences in prevalence rates obtained by a comparison of the average direct question response across groups are biased due to sensitivity. In such cases, a common practice is to estimate the difference in prevalence rate via a regression of the list experiment response on treatment, and indicator for group membership, and the interaction between the treatment and group membership indicators. The coefficient on the interaction term is an estimate of the difference-in-prevalence rates.

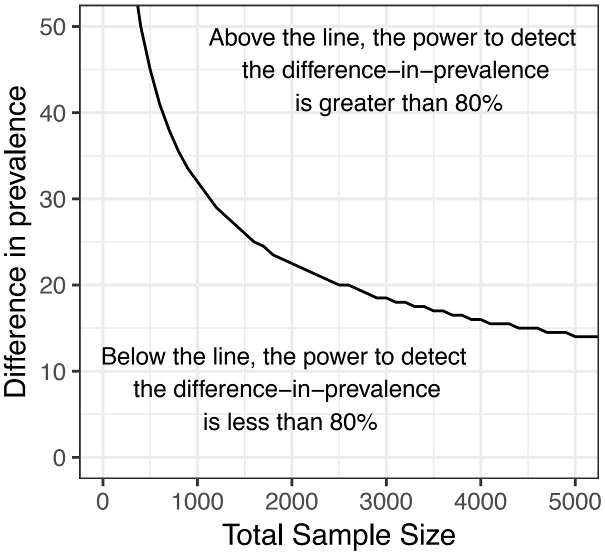


Figure 2: Power to Detect Differences in Prevalence Rates between Groups is Low Except When the Difference or the Sample Size is Very Large. On the y-axis is the difference in prevalence rates between the two groups, which determines the power to detect the difference. The power for this design does not depend on the prevalence rate itself.

As described in Samii (2012), the high variance of the list experiment frustrates the comparison of prevalence rates across groups, regardless of whether those groups are formed experimentally or on the basis of background attributes. Figure 2 shows that the power to detect even substantial differences in prevalence rates is abysmal. Differences must exceed 25 percentage points before a 2,000 unit study has 80% power to detect them; they must be 20 points or more in the case of a 3,000 unit sample. Conclusively demonstrating that two groups have different prevalence rates requires extreme differences and very large samples.

Improving the Power of the List Experiment Design

The high variance of the list experiment technology has not escaped the notice of survey methodologists, who have generated a suite of improvements over the standard list experiment design, most of which have variance reduction as their primary goal. In this section, we describe each innovation in terms of the effective sample size improvement over the standard design, allowing a direct comparison of designs using a common metric.

Droitcour et al. (1991) proposes the double list experiment in which all subjects participate in two list experiments with different control items but the same sensitive item. Subjects are randomly assigned to see the treatment list in one experiment but not the other; the combined estimate from each experiment has approximately 50% the variability of the equivalent single list experiment. Glynn (2013) focuses on the selection of the control items and suggests that researchers choose control items that are negatively correlated with one another, with the goal of reducing the variance of the control items as much as possible. If the variance of the control items were equal to zero, the list experiment would be exactly as precise as a direct question conducted on half the sample. In such a scenario, it must be mentioned, the list experiment would provide no cover at all and those with the sensitive trait in the treatment group would be exactly identified. In practice, however, the standard deviation of control items is difficult to reduce much below 0.65.

Other scholars have proposed various methods for combining list experiments with other sources of information. Blair et al. (2014) proposes a combined list and endorsement experiment that succeeded in reducing the variance of the list experiment by 12%. Aronow et al. (2015) derive a method for combining list and direct questions by conducting a list experiment among those subjects who do not directly admit to the sensitive trait. This procedure recovers the precision of the direct question among those whose answers are presumed to be truthful and combines it with the unbiased estimate generated by the list experiment among those who do not admit. In their applications, the combined estimator decreased variance by 12% to 50%. Chou et al. (2018) provide a generalization of Aronow et al. (2015) to any subgroup among whom the true prevalence rate is known. Auxiliary information of this sort, though rare, can dramatically reduce variability. In their application to support for an

	% Variance Reduction	Effective Sample Size
Double list experiment (Droitcour et al. 1991)	50%	4,000
Control item advice (Glynn 2013)	40% - 55%	3,400 - 4,440
Combined list and endorsement experiment design (Blair et al. 2014)	12%	2,250
Combined list experiment and direct question design (Aronow et al. 2015)	12% - 50%	2,250 - 4,000
Using auxiliary information (Chou et al. 2018)	88%	25,000

Table 1: Variance reduction and increase in effective sample size relative to a N = 2,000 standard design for alternative list experiment designs.

antiabortion ballot measure, auxiliary information in the form of known vote totals reduced the variance of the list experiment by an estimated 88%.

Model-based methods to improve power of the list experiment include the linear regression, non-linear least squares, and maximum-likelihood models proposed in Imai (2011). Maximum likelihood models have also been proposed for the LISTIT design (Corstange 2009; Blair and Imai 2012).[2] Subsequent modifications are designed to accommodate violations of the no liars assumption including ceiling and floor effects (Blair and Imai 2012) and nonstrategic misreporting due to satisficing (Blair et al. Forthcoming).

Table 1 shows how each of these methods help to decrease variance. The final column shows what the sample size of the standard list experiment design would need to be in order to achieve the same precision as each method conducted on a sample of 2,000 subjects. The feasibility of each improvement will vary depending on the application; sometimes unavoidable features of the setting will cause violations of the specific assumptions invoked by each design. For example, if sensitivity bias affects different subgroups in opposite directions, the required assumption of monotonicity in Aronow et al. (2015) would be violated. If the sensitive trait is a behavior, it can be difficult to construct an endorsement experiment – in such cases, the proposal in Blair et al. (2014) would not be feasible.

[2] The LISTIT design likely also increases precision relative to the conventional design, but we were unable to include it in Table 1 because we lack empirical estimates of the extent of variance reduction.

5. Assessing the Assumptions of the List Experiment

We rely on the list experiment estimate of the prevalence of the sensitive item as a measure of the true prevalence rate. To do so, we invoke the four assumptions described above. Before we conduct the meta-analysis, we assess the validity of these assumptions. In particular, we examine the key no design effects assumption.

For each list experiment we collected, we perform a test of the no design effects assumption from Blair and Imai (2012), described above. Out of 166 list experiments for which we have sufficient information to conduct the test, only 6 fail. The test itself may be underpowered to detect moderate violations of no design effects, but the overwhelming success rate suggests at a minimum that large violations of no design effects are implausible. We include the few studies that fail for fear they are false positives; our conclusions are robust to the inclusion or exclusion of these few studies.

6. Sensitivity Bias by Research Area

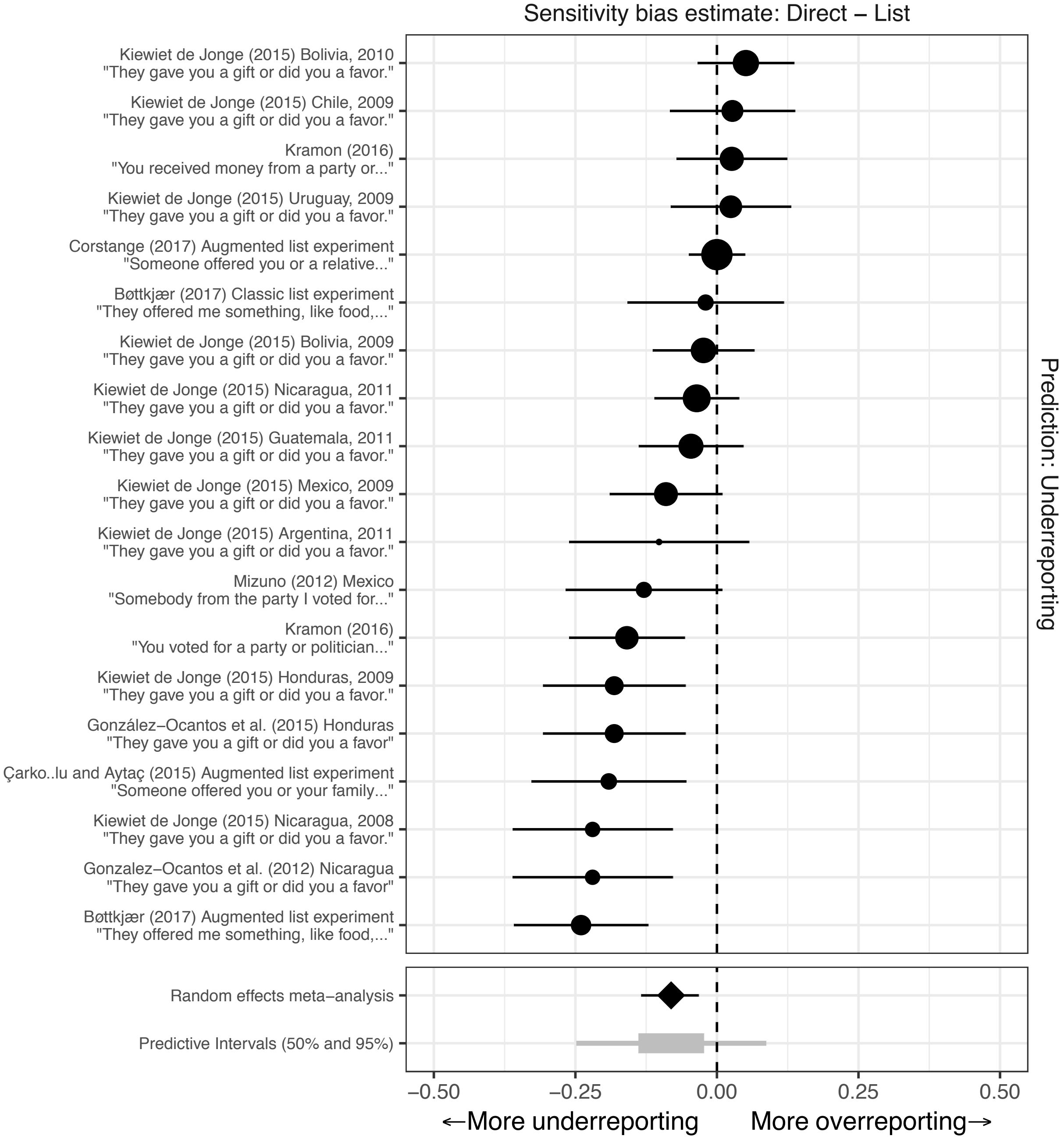


Figure 3: Estimates of Sensitivity Bias for Vote Buying

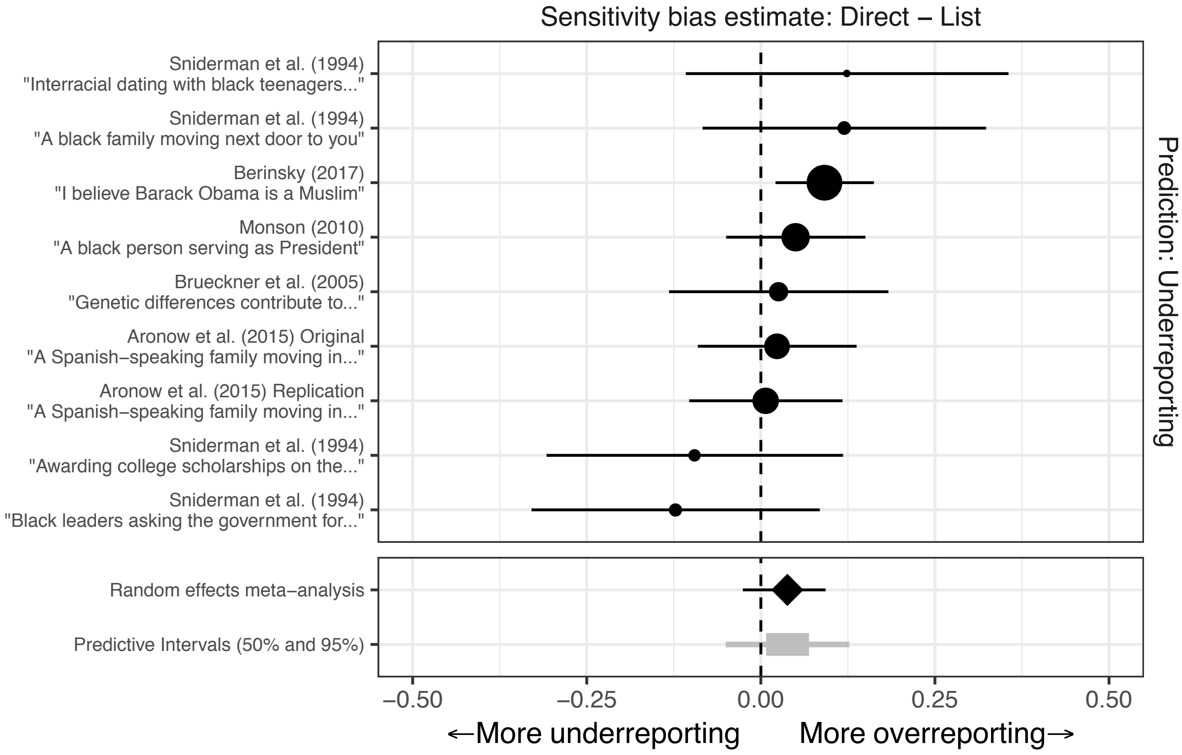
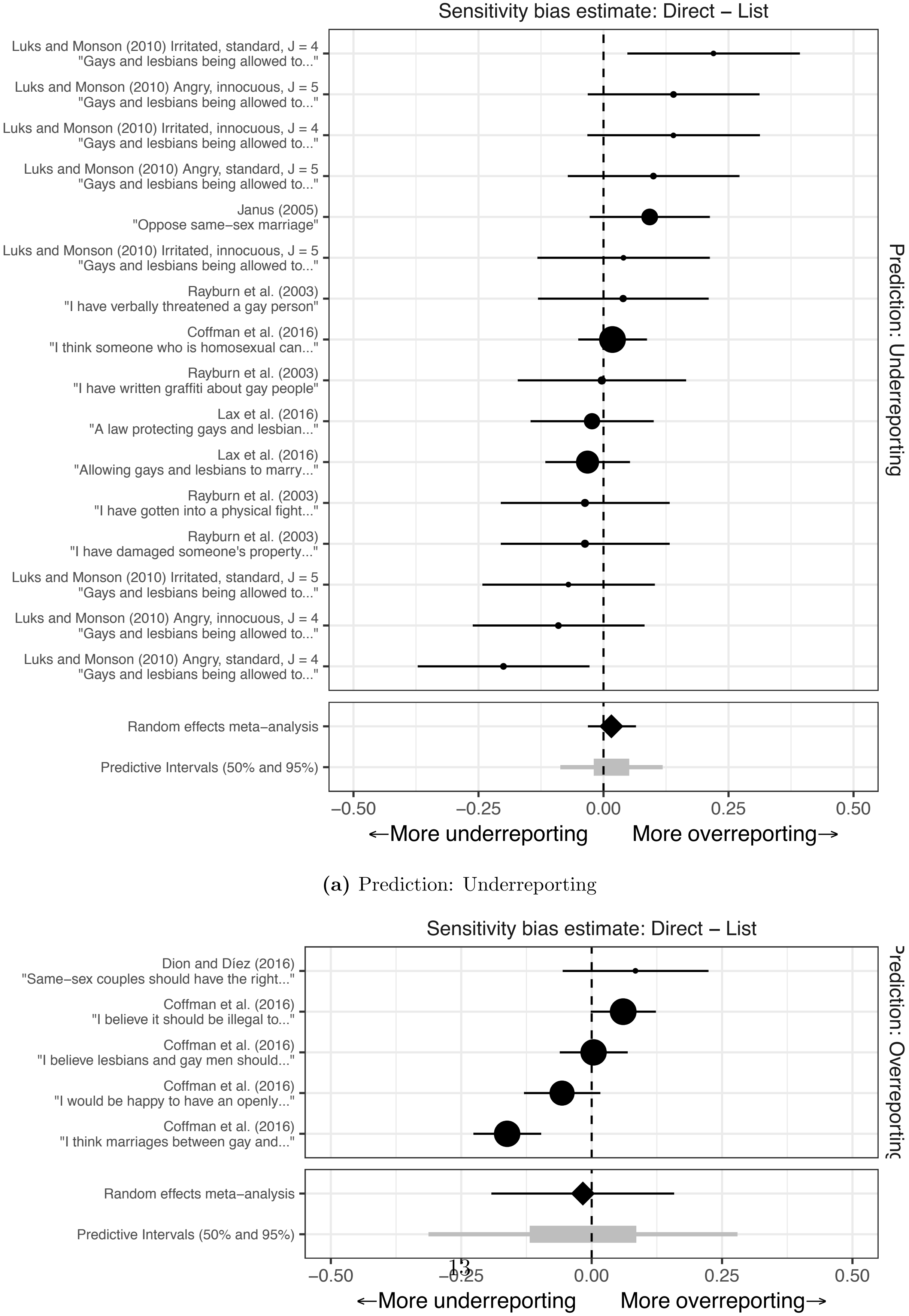


Figure 4: Estimates of Sensitivity Bias for Racial Prejudice



(a) Prediction: Underreporting

(b) Prediction: Overreporting

Figure 5: Estimates of Sensitivity Bias for Sexual Orientation Prejudice

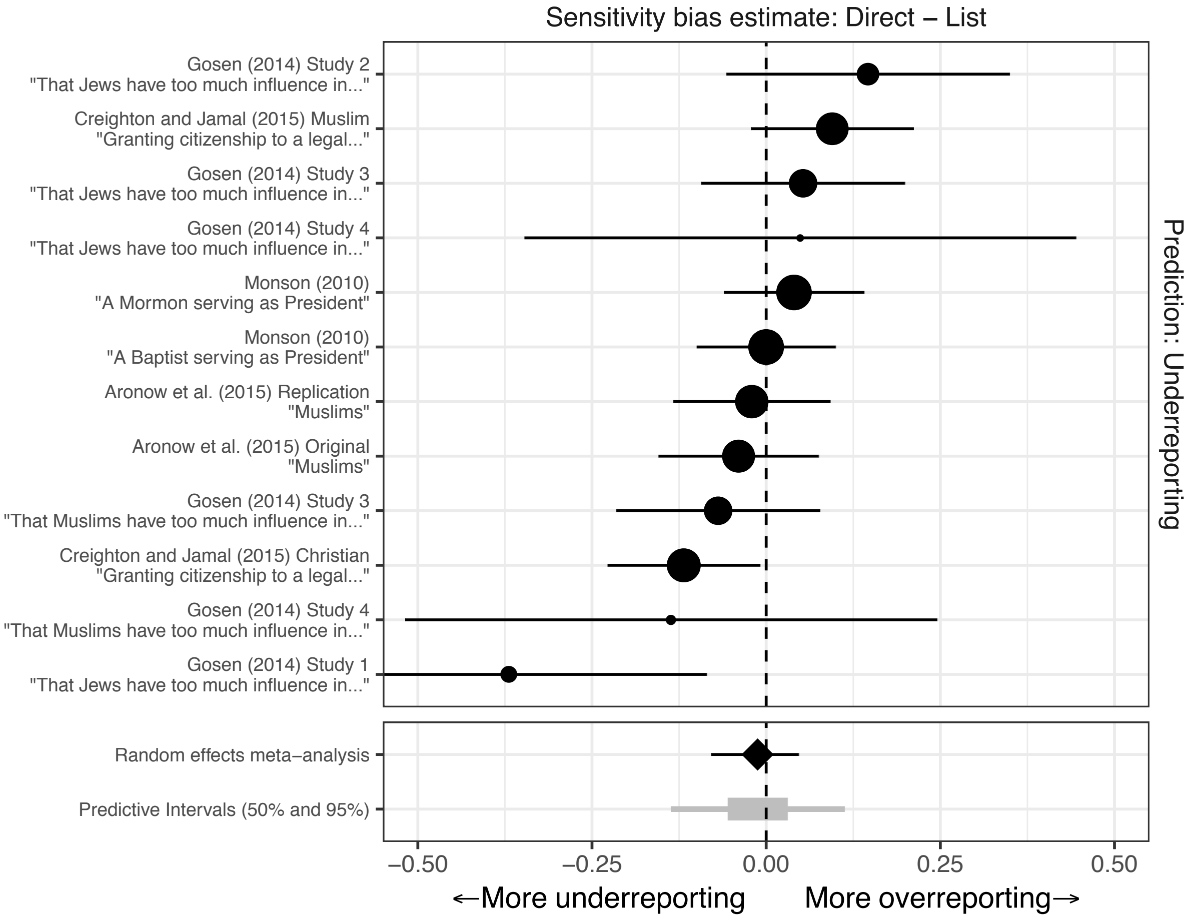


Figure 6: Estimates of Sensitivity Bias for Religious Prejudice

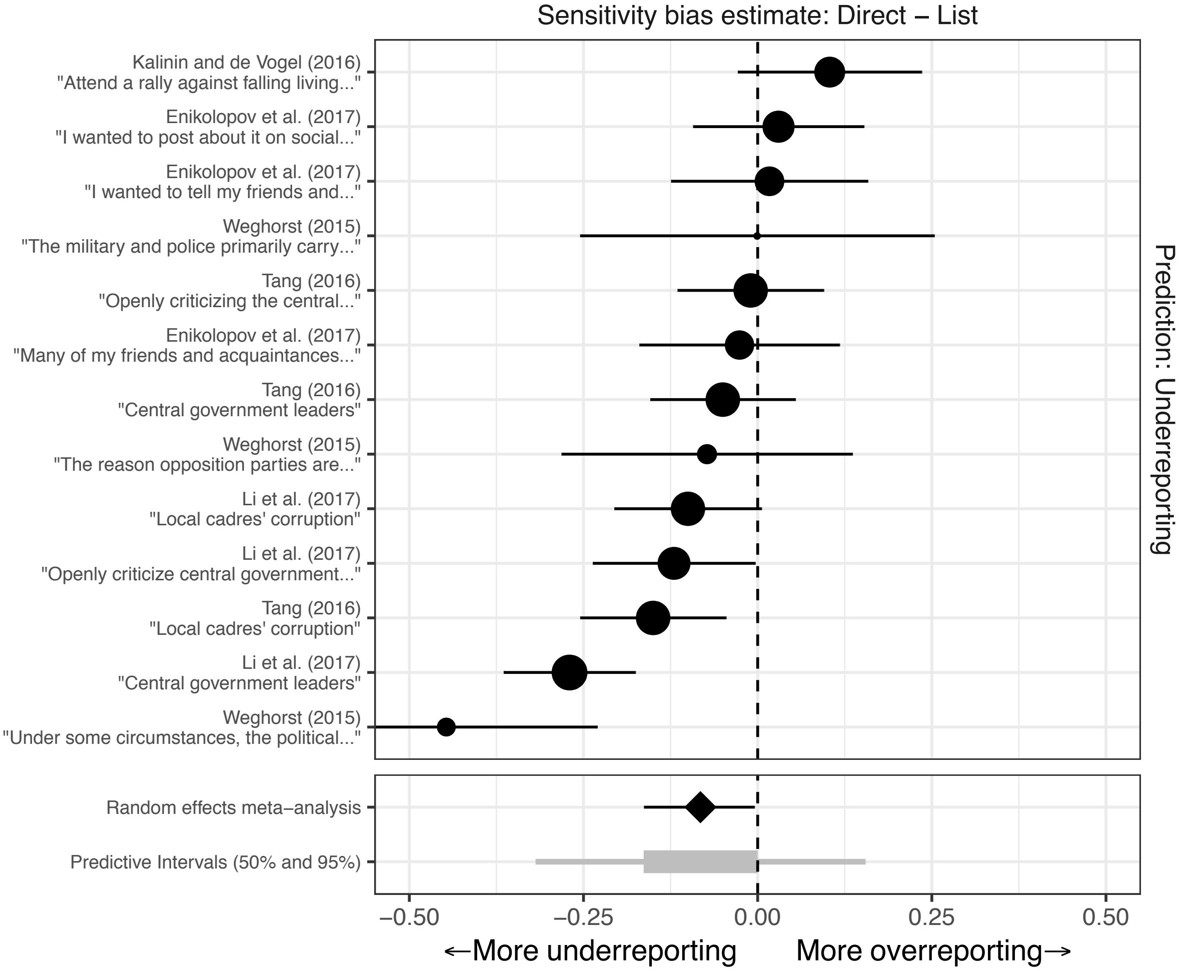


Figure 7: Estimates of Sensitivity Bias for Political Attitudes in Authoritarian Regimes

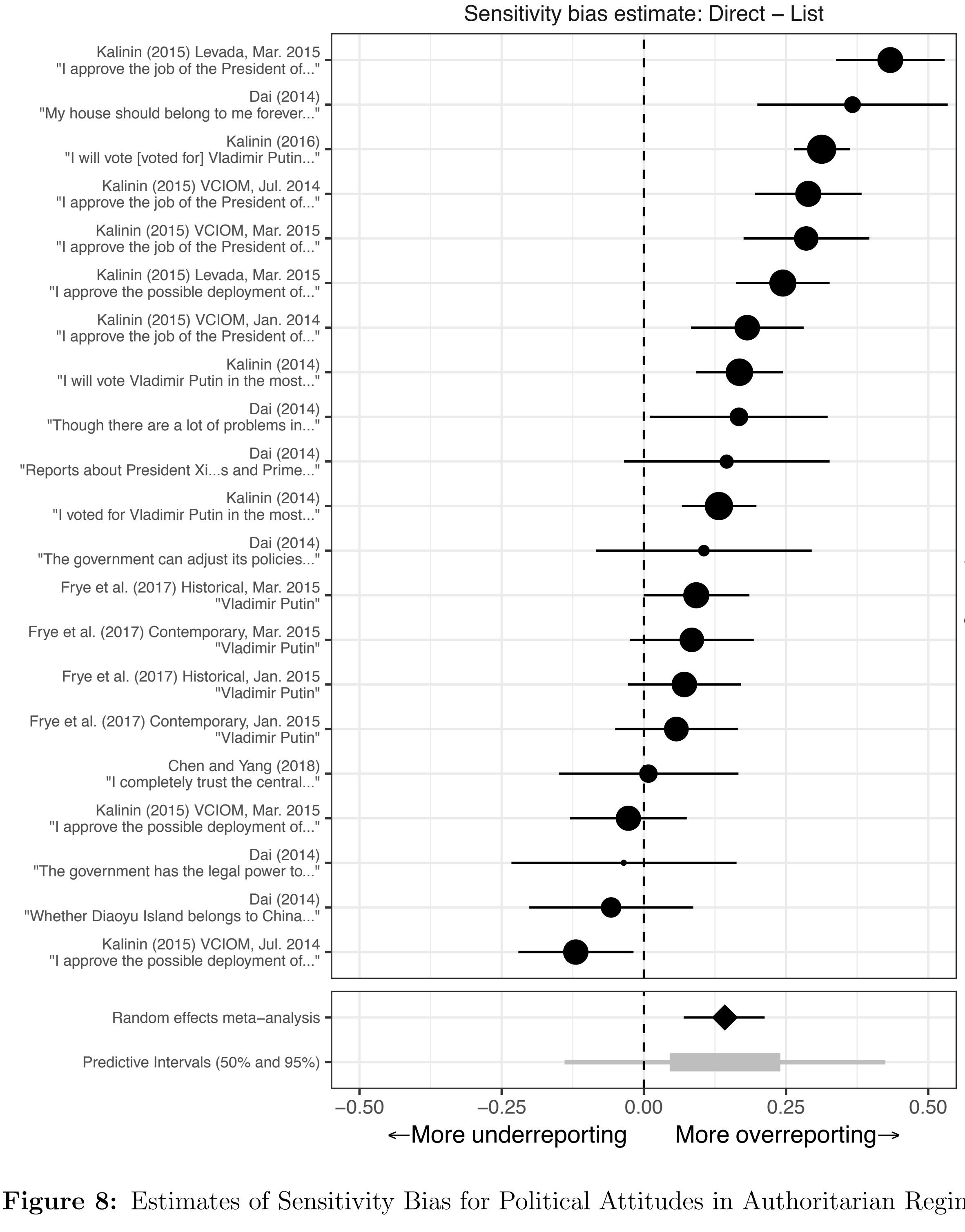


Figure 8: Estimates of Sensitivity Bias for Political Attitudes in Authoritarian Regimes

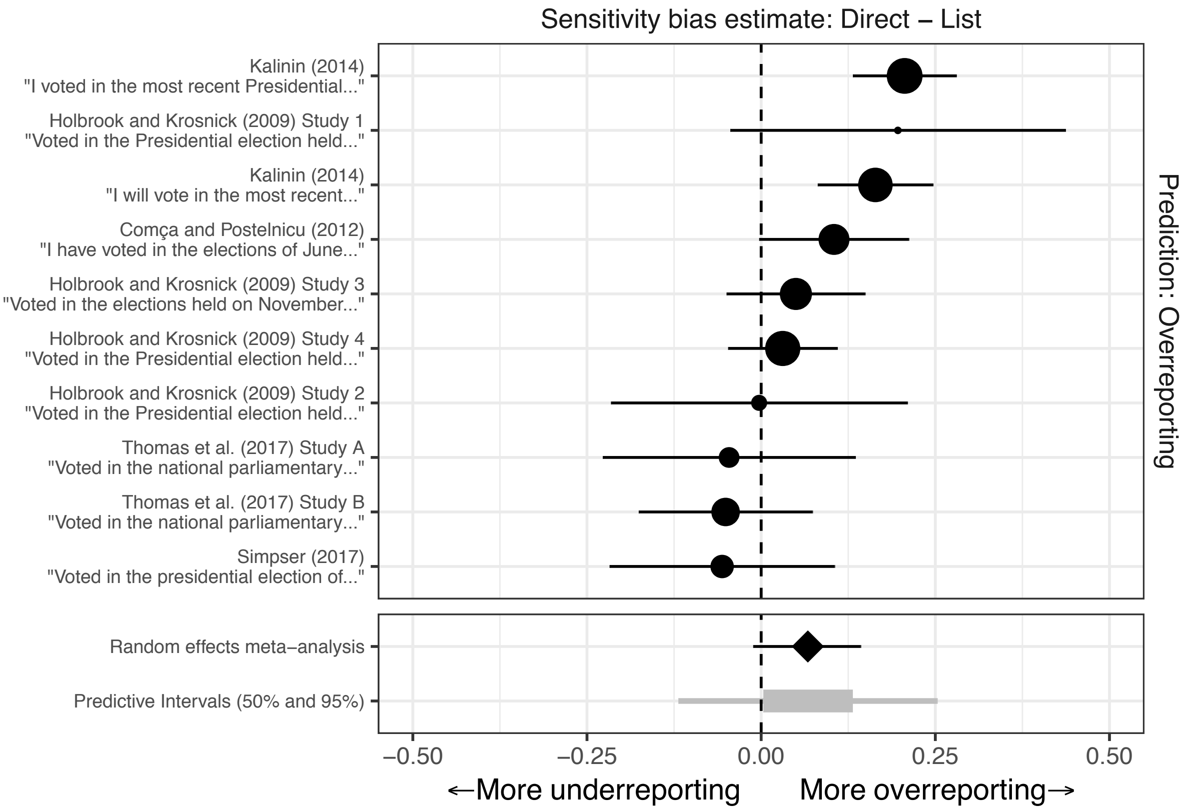


Figure 9: Estimates of Sensitivity Bias for Turnout

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.

Blair, Graeme and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1):47–77.

Blair, Graeme, Kosuke Imai and Jason Lyall. 2014. "Comparing and combining list and endorsement experiments: Evidence from Afghanistan." *American Journal of Political Science* 58(4):1043–1063.

Blair, Graeme, Winston Chou and Kosuke Imai. Forthcoming. "List Experiments with Measurement Error." *Political Analysis* .

Chou, Winston, Kosuke Imai and Bryn Rosenfeld. 2018. "Sensitive Survey Questions with Auxiliary Information." *Sociological Methods & Research* . Forthcoming.

Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT." *Political Analysis* 17(1):45–63.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher and Trena M. Ezzati. 1991. The Item Count Technique as a Method of Indirect Questioning: a Review of its Development and a Case Study Application. In *Measurement Errors in Surveys*, ed. Biemer, Groves, Lyberg, Mathiowetz and Sudman. John Wiley & Sons, chapter 11, pp. 185–210.

Flavin, Patrick and Michael Keane. 2009. "How angry am I? Let me count the ways: Question format bias in list experiments." Working paper, Department of Political Science, Baylor University.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York: W.W. Norton.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1):159–172.

Imai, Kosuke. 2011. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association* 106(494):407–416.

Janus, Alexander L. 2010. "The Influence of Social Desirability Pressures on Expressed Immigration Attitudes." *Social Science Quarterly* 91(4):928–946.

Kramon, Eric. 2016. "Where is vote buying effective? Evidence from a list experiment in Kenya." *Electoral Studies* 44:397–408.

Samii, Cyrus. 2012. "List Experiments as Outcome Measures." Unpublished research note, Department of Politics, New York University.

Zigerell, L. J. 2011. "You Wouldn't Like Me When I'm Angry: List Experiment Misreporting." *Social Science Quarterly* 92(2):552–562.