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

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

Forthcoming, *American Political Science Review*

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. We make three contributions. First, we propose a social reference theory of sensitivity bias to structure expectations about survey responses on sensitive topics. Second, we explore the bias-variance tradeoff inherent in the choice between direct and indirect measurement technologies. Third, to estimate the extent of sensitivity bias, we meta-analyze the set of published and unpublished list experiments (a.k.a., the item count technique) conducted to date and compare the results with direct questions. We find that sensitivity biases are typically smaller than 10 percentage points and in some domains are approximately zero.

[‡]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. Ever since, 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, as discovered in the earliest National Election Studies, survey respondents overreport voter turnout by 10 to 15 percentage points (Campbell et al. 1960). In the other direction, one meta-analysis found that 30-70% of clinically-confirmed drug users underreport drug use (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 leaves open to interpretation *who* desires a particular response and why a respondent would care. Second, it does not capture other sources of bias beyond conformity with perceived social norms, including monetary costs and threats to personal safety. In this paper, we build on frameworks from sociology, 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, with approaches falling into three broad categories. First, survey researchers can change the form of the question with techniques like projective questioning (Haire 1950) or the method studied in this paper, the list experiment, also known as the unmatched count technique and originally called the item count technique (Miller 1984). Second, they can change the social context in which the question is asked (Hatchett and Schuman 1975; Silver et al. 1986). Third, they can attempt to identify which types of people are most prone to giving false answers (Snyder 1987; 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 likely to be a problem and if it is, how best to address it.

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 (in part or in full) 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 (Warner 1965). 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), the block total response technique (Raghavarao and Federer 1979), and the cross-wise technique (Yu et al. 2008) 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 first identified in Miller (1984) for the choice between direct questions and list experiments. The broad strokes of our argument apply to these other formats as well. For a comparison of the bias and variance of these methods, see Rosenfeld et al. (2016).

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 major interest to political scientists including 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 substantive domains. We find evidence of overreporting of support for authoritarian regimes and suggestive evidence of overreporting for voter turnout, underreporting in vote buying and criticism of authoritarian regimes, but nearly no evidence of sensitivity bias in measures of prejudice.

Determining whether sensitivity bias is likely to be 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. Even when we can be confident that list experiments generate unbiased prevalence estimates, they are far less accurate (in terms of mean-squared error) than direct questions at typical sample sizes.

A Social Reference Theory of Sensitivity Bias

Why do questions about sensitive topics in surveys yield 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).

To formalize sensitivity bias, we assume that a subject i harbors a latent true value of the sensitive trait, D_i^*. The latent value may differ from the response that a subject would give if asked directly, D_i. We assume that D_i^* exists 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 ($D_i = \texttt{NA}$). 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, respondent memory failures, 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, which could result from either conscious or unconscious processes (Tourangeau and Yan 2007). When D_i^* is binary (the case we consider here), 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.[1] The difference $D_i - D_i^*$ may vary by unit or subgroup, and sensitivity bias can be calculated by subgroup by taking the two expectations over a subset of units.

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.

The dominant explanation for sensitivity bias since the 1950s has been social desirability bias (Maccoby and Maccoby 1954). According to Fisher (1993, p. 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

---

[1]The magnitude of δ can also be thought of as function of the prevalence rate π^* and the rate of false reporting. In the case of overreporting, δ is the product of $1 - \pi^*$ and the false reporting rate among those for whom $D_i^* = 0$; for underreporting, δ is the product of π^* and the false reporting rate among those for whom $D_i^* = 1$.

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 sensitivity bias and have shown that different people withhold at different rates (Snyder 1987; Paulhus 1991; Berinsky 2004). We interpret these findings as evidence of individuals' idiosyncratic beliefs about relevant social referents. 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.

Perhaps the most salient social referent for subjects is the interviewer asking the survey questions (Feldman et al. 1951). Subjects may presuppose (rightly or wrongly) that interviewers have an opinion about what the correct attitude to hold is. Using randomized experiments, scholars have demonstrated interviewer effects 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 (Corstange 2014); 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 (Lyall et al. 2013).

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 (Tourangeau et al. 2000). For taboo topics, the act of responding, separate from the content of the response, may itself be sensitive. In such cases, sensitivity bias may manifest as nonresponse. Third, respondents may fear their responses will be disclosed to authorities such as governments, criminals, armed groups, or employers. The worry here 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 for a given respondent if and only if 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 themself.

2. A respondent perception that the social referent can infer the subject's response to the sensitive question either exactly or approximately.[2]

3. A respondent perception about what response (or nonresponse) the social referent prefers.

4. A respondent perception that failing to provide the response preferred by the social

---

[2]For a formalization of the idea that respondents may still misreport even if social referents can only partially learn their responses, see Simpser (2017).

referent would entail costs to themself, other individuals, or groups. Costs may be social (embarrassment), monetary (fines), or physical (jail time or personal violence).

These four elements can help structure expectations about when to worry about sensitivity bias. These expectations, importantly, may vary across different types of respondents (for example, in some circles racist views may be something to hide, but in others something to flaunt). If respondents do not have social referents in mind, or if they truly believe their answers will remain confidential, or if they do not know the preferences of the referent, or if they perceive the referent to be powerless to inflict costs, then standard survey measures may be free of sensitivity bias. When all four elements are present, articulating how they play out in a specific context can generate educated guesses about the plausible direction and magnitude of bias. These guesses can be tested empirically by collecting survey and focus group data from respondents about their second order beliefs about relevant reference groups or directly estimating sensitivity bias.

This theory helps to distinguish sensitivity bias from other forms of measurement error, such as the distinction between self-deception and recall failures (Holbrook and Krosnick 2010). 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 election, direct questions will be biased, but not because of sensitivity. The theory also helps us to understand how the same question may be sensitive in one context but not another. Respondents may perceive that different referents prefer different answers, which raises the possibility of cross-cutting sources of bias. Related to this point is the idea that sensitivity bias will depend on survey mode. The important social referents in a face-to-face household survey will likely differ from those in an online or telephone survey (Hochstim 1967).

A voluminous literature examines the link between implicit and explicit attitudes (Green-

wald and Banaji 1995; Greenwald et al. 1998). Since implicit attitudes are thought to operate at an unconscious level, respondents themselves may be unable to self-report them accurately. As a result, our theory likely applies to explicit attitudes and behaviors only (see also 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 identified 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 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). One important behavior that characterizes clientelistic systems, vote buying, 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 (Weitz-Shapiro 2012).

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 and Romania, Mares and Young 2019). In some contexts, however, voters may not be

Social referent	Respondent beliefs		
	Referent can learn about 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	Legal sanction
Neighbors	Possibly, depending on anonymity protections	No	Self presentation
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	State repression
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.

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 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" (p. 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, p. 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 (Mares

et al. 2016) or to "acknowledge that the handout influenced their vote" (Brusco et al. 2004, p. 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 (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 relevant 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 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 (Powell 2013). 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 (Kane et al. 2004).

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. 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.[3] As a result, dictators may exert pressure on citizens to profess higher levels of support for the regime than they truly hold (a phenomenon labeled "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 suggests this is exactly what is happening. High-quality surveys from Russia's only independent polling agency (Yuri Levada Analytical Center 2019) estimate Russian support for President Vladimir Putin at 89 percent. According to the 2012-13 World Values Survey China, support for the central government in China is a full 93 percent (Tang 2016). The four sensitivity bias criteria are plausibly met in these settings: (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

---

[3]A similar problem is faced by opposition supporters, whose risky choice to protest or otherwise dissent may depend on the difficult-to-estimate level of public support for their cause. Indeed, in some accounts of the fall of the Berlin Wall, revolution only occurred following small protests in Leipzig, which revealed to East Germans their shared antipathy to the regime and willingness to act for change (Lohmann 1994).

may include harassment, imprisonment, or worse (Frye et al. 2017; Chen and Yang 2018).

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% (this pattern persists, see Vavreck 2007). Recent work has shown that this problem affects voter participation research in developing democracies as well (Adida et al. 2019). 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 a particular election. 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 that they did not vote. Because whether or not a person has voted is a public record in many places, 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.[4]

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.[5] Subjects are randomly assigned to treatment and control groups. Those in the control group report how many of a list of control items apply to them; treatment group subjects report how many of a list of the control items plus an additional sensitive item apply to them. The prevalence rate of the sensitive item can be estimated by calculating the difference in the average response given by each group.

This technique mitigates sensitivity bias by hiding the sensitive item response from many possible social referents, including the interviewer, bystanders, or later data consumers, except in the case of floor and ceiling responses (Miller 1984; Kuklinski et al. 1997). The list experiment does not address all sources of sensitivity bias. Intrusive questions may still generate bias because the text of the question includes the same sensitive item text as the direct question. Further, because the list experiment does not hide answers from the respondents themselves, the method will not reduce sensitivity biases arising from self-image concerns.

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 during the election. To do so, the authors use a list experiment and a

---

[4]Avoiding such embarrassment is one hypothesized mechanism behind the social pressure get-out-the-vote treatments (Gerber et al. 2008).

[5]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 (see also Imai 2011).

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 and whether a bribe influenced the respondent's vote from the 2007 Kenya post-election survey reported in Kramon (2016).

direct question. Respondents were randomized into a control group or one of 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.

In the "Received" treatment group, the same script was read but with a fifth item added to the list:[6]

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

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

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

---

[6]Typically, item order is randomized and the sensitive item is not necessarily the last item.
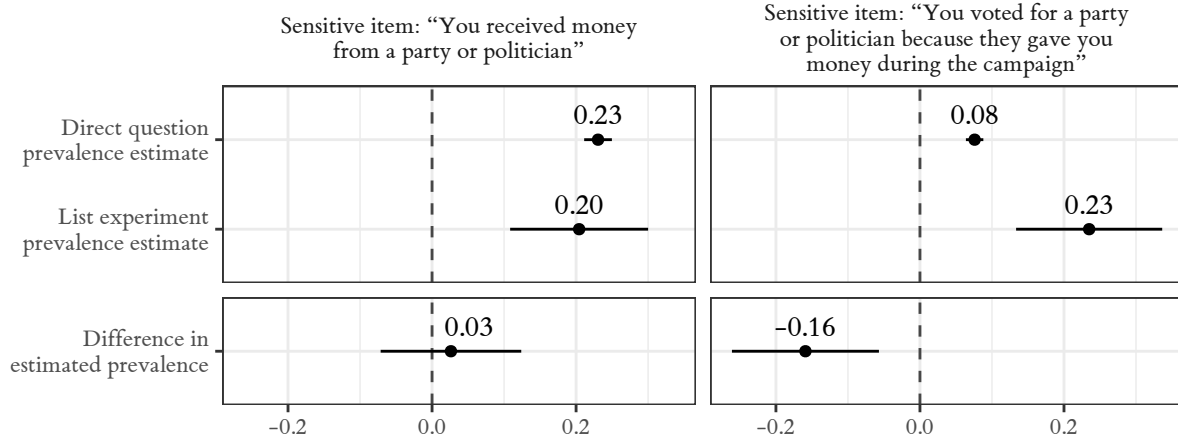
Figure 1: Estimated prevalence of vote buying from the list experiments and the direct questions for two sensitive items presented in Kramon (2016): whether the respondent received a bribe and whether a bribe influenced the respondent's vote.

Using data from the Kramon (2016) postelection survey in Kenya (shown in Table 2), 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 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 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.[7] Under their identifying assumptions, list experiments are unbiased, but high variance.

Consider a study of N subjects with a true prevalence rate (π^*) in which the direct question has a sensitivity bias (δ). 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

---

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

of the direct question estimator $(\widehat{\pi})$ is given by

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

The variance of the list experiment estimator $\widehat{\pi^*}$ under a balanced design (in which exactly half of the units are assigned to the treatment list) is given by:

$$\mathbb{V}(\widehat{\pi^*}) = \frac{1}{N-1}\left\{\pi^*(1-\pi^*) + 4\mathbb{V}(Y_i(0)) + 4\text{cov}(Y_i(0), D_i^*)\right\} \tag{2}$$

where $\mathbb{V}(Y_i(0))$ is the variance of the control item count and $\text{cov}(Y_i(0), D_i^*)$ is the covariance between the control item count and the true sensitive item response. See the supplementary materials for derivations of these variance expressions.

Plugging in $N = 2{,}000$, $\pi^* = 0.5$, and $\delta = 0.1$ and taking the square root yields a standard error of the direct question estimator of 0.0111, or 1.11 percentage points. Using the average values of the variance of the control items (0.075) and their covariance with D_i^* (0.025) estimated from the set of studies in our meta analysis, we obtain a standard error for the list experiment of 0.0416, or 4.16 percentage points.[8] For the same number of subjects, the list experiment is $(4.16/1.11)^2 \approx 14$ times more variable than the direct question. Stated differently, a researcher would need a sample of 28,000 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 item and their responses are further obscured by adding noise.

This bias-variance tradeoff interacts with the goal of the research. We identify four main goals: estimating a prevalence rate, demonstrating the presence of sensitivity bias, estimating the difference in prevalence rates across groups, and estimating the difference in sensitivity

---

[8]In the replication archive, we provide DeclareDesign code (Blair et al. 2019) to simulate a list experiment design with these parameters.

bias across groups.[9]

When the primary research goal is obtaining a good estimate of the prevalence rate of a sensitive item (e.g., Gervais and Najle 2018, who 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 two important 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 5.5 points to prefer a list experiment; at 2,000, the bias must be greater than 4 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 sensitivity bias for same-sex marriage could not be distinguished from zero. The main 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 exceed 16 percentage points in order to reach 80% power; even at a sample size of 2,000, power to

---

[9]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).
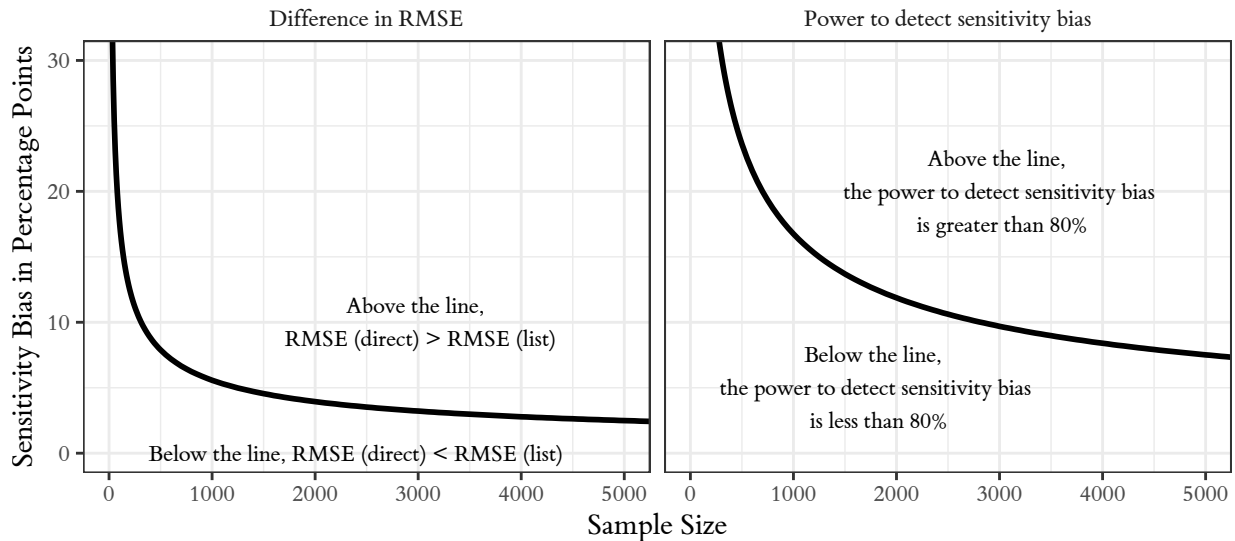
Figure 2: For a Given Sample Size, Whether the List Experiment is Preferred to Direct Questions Depends on the Expected Level of Sensitivity Bias. For designs that aim to estimate the prevalence rate of the sensitive item (left panel), the root mean-squared error indifference curve characterizes the tradeoff. The curve in the right panel indicates the bias-sample size combinations at which the power to detect sensitivity bias is 80%.

detect biases of 10 percentage points is well below the conventional power target.

Many social scientific theories predict that prevalence rates will differ according to sub-groups 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. Interestingly, for a comparison of direct question responses to generate biased estimates of the difference in prevalence rates across groups, sensitivity must affect responses in each group differently, since otherwise, the biases would cancel each other out when the group prevalence rates are differenced. In order for sensitivity to vary across groups, respondent beliefs about social referents – their preferences, their ability to monitor, or the costs they impose – must be heterogeneous.

In such cases, researchers can employ a list experiment to measure outcomes, then estimate the difference in prevalence rates using an interaction model. As described in Samii

(2012), however, this estimator has extremely high variance. Figure 3 panel (a) 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 using list experiments requires extreme differences and very large samples.

Some studies hypothesize not differences in the prevalence rates across sub-groups but rather differences in sensitivity bias. For example, Lax et al. (2016) study "whether social desirability effects differ by a respondent's partisan identification and religious affiliation," and Kiewiet de Jonge (2015) studies how sensitivity bias varies by income level. As panel (b) of Figure 3 shows, such tests are also quite sample hungry.

(a) Difference-in-prevalance
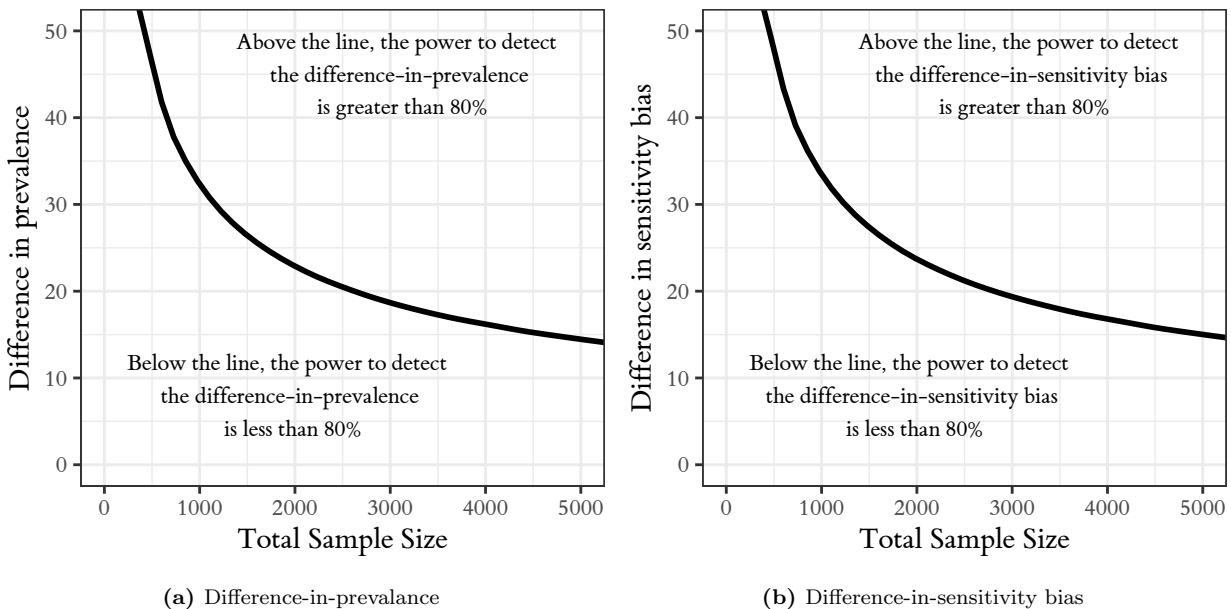
(b) Difference-in-sensitivity bias

Figure 3: Power to Detect Differences in Prevalence Rates or Differences in Sensitivity Bias between Groups is Low Except When the Difference or the Sample Size is Very Large. On the y-axes is the difference in either the prevalence rate or the sensitivity bias between the two groups, which affects the power to detect the difference.

Improving the Power of the List Experiment Design

After introducing the technique, Judith Droitcour Miller wrote in her 1984 dissertation that "the variance of the item-count estimator is potentially high" (p. 67). She noted that the loss of power comes from aggregating the sensitive item with other items and measuring only the count, and the fact that "reports of deviant behavior are collected from only half of the total sample" (p. 57). She laid out a series of practical suggestions for constructing the lists and alternative designs to reduce the variance. Many of these ideas have been taken up by survey methodologists in the decades since. In this section, we describe each modification in terms of the effective sample size improvement over the standard design, allowing a direct comparison of designs using a common metric.[10]

Just after introducing the standard design, Miller (1984) proposed the double list experiment design, in which all subjects participate in two list experiments with different control items but the same sensitive item (see also Droitcour et al. 1991; Glynn 2013). Subjects are randomly assigned to see the treatment list in one experiment but not the other; the combined estimate has approximately 50% the variability of the equivalent single list experiment.

Miller (1984) offers detailed guidance on selecting control items to reduce the variance of the list experiment. These include trading off the length of the lists with privacy protection, selecting some low- and some high-prevalence items to avoid many respondents answering at the floor ("no" to all items) or ceiling ("yes" to all items), inducing negative correlation across items within the list, and inducing positive correlation between the two lists in a double list design (pp. 68-70). These list design tips are further explored in more recent methodological work (Blair and Imai 2012; Glynn 2013).

Other scholars have introduced methods for combining list experiments with other sources

---

[10]For a parallel comparison of design variants for the randomized response technique, see Blair et al. (2015).

	% Variance Reduction	% Increase in Effective Sample Size
Double list experiment (Miller 1984)	50%	100%
Control item advice (Miller 1984; Glynn 2013)	40% - 55%	70% - 122%
Combined list and endorsement experiment design (Blair et al. 2014)	12%	14%
Combined list experiment and direct question design (Aronow et al. 2015)	12% - 50%	14% - 100%
Using auxiliary information (Chou et al. 2018)	88%	733%

Table 3: Variance reductions and increases in effective sample size from alternative list experiment designs.

of information. Blair et al. (2014) proposes a combined list and endorsement experiment that succeeded in reducing variance 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. 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. In their application to support for an antiabortion ballot measure, auxiliary information in the form of known vote totals reduced the variance of the list experiment by 88%.

Model-based methods to reduce the variance 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).[11] Table 3 shows how each of these methods help to decrease variance and increase effective sample sizes. 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.

Table 3 can be used in conjunction with Figures 2 or 3 to guide sample size decisions for list experiments that employ these alternative designs. At a given nominal sample size, we

---

[11]The LISTIT design likely also increases precision relative to the conventional design, but we were unable to include it in Table 3 because we lack empirical estimates of the extent of variance reduction.

can look up what the effective sample size will be, then reference the figures to determine if a list experiment is an appropriate measurement tool for a given research scenario.

Meta-Analysis Research Design

In this section, we present a meta-analysis of list experiments to characterize the level of sensitivity bias in four 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.[12] We use this evidence to assess where each research area falls on the bias-variance tradeoff. These estimates can be used to help researchers make context-specific decisions about survey question designs.

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, colleagues have told us of list experiments that were never written up and whose data are long since lost. We searched Google Scholar, SSRN, Dataverse, and seven years of political science conference programs with the search terms "list experiment," "item count technique," and "unmatched count technique," and their abbreviations. In addition, we traced citations from several influential early papers on list experiments including Judith Droitcour Miller's dissertation which introduced the technique and from each of the studies identified in our keyword searches. 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 topics.

We gathered statistical information about the list experiments using a cascading data

---

[12]For a contemporaneous meta-analysis in sociology that investigates the relationship between design features of the list experiment and sensitivity bias, see Li and den Noortgate (Forthcoming).

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 crosstabs of list experiment responses by treatment condition (similar to Table 2). For those studies, we calculated the difference-in-means estimate of prevalence and its standard error. Finally, if neither the data nor the crosstab were available, we searched the paper for the estimated prevalence rate and its standard error. In the rare case that 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 concern 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 and so chose not to pursue publication. Our main approach to this problem was to energetically seek out published and unpublished studies. Nevertheless, we are sure that we have missed some studies that were conducted but for whatever reason were never written up. Reasoning about the impact of selection into our meta-analysis is complex for at least two reasons. First, many studies had multiple inferential goals (estimating a prevalence rate, demonstrating the existence of sensitivity bias, measuring group differences in prevalence rates, or some combination of these) and it was not obvious which, if any, was the primary driver of publication incentives. For this reason, we do not present diagnostics such as funnel plots or p-curves. Funnel plots require a common quantity of interest thought to be subject to the publication filter and p-curves require a common theoretical null distribution of p-values. Secondly, it is not always clear which result – null or significant – is theoretically congenial in a given setting. Some theories predict the absence of sensitivity bias (e.g., Coppock 2017) while others predict its presence (e.g., Gonzalez-Ocantos et al. 2012), so intuiting how the file-drawer problem may distort our meta-analysis is challenging.

The direct question estimates of prevalence all come from the original authors. The

24

majority of studies asked the direct question to either their entire sample or a random subset; three of our studies referred to a direct question estimate obtained by others. We logged whichever direct question estimate was reported by the original authors. We elected not to independently obtain direct question 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 for direct question estimates introduces a second source of selection in addition to publication bias.

Our measure of sensitivity bias is the difference between the list and direct estimates. We estimated the standard error of the difference as $\text{SE}(\text{difference}) = \sqrt{\text{SE}(\text{list})^2 + \text{SE}(\text{direct})^2}$. 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 using data from the same respondents. 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, then by 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.[13]

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 and δ represents the true sensitivity bias for a given study. We model bias as $\delta \sim \mathcal{N}(\mu, \tau)$. The parameters of this distribution are μ, the grand mean of sensitivity bias and τ, the standard deviation of true effect sizes. We conduct Bayesian estimation using Stan (Carpenter et al. 2017), adopting the default improper uniform priors for δ, 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 true level of bias could 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

---

[13]We also coded the mode of survey administration for each study. Subjects may have different social referents in mind depending on whether the survey was conducted face-to-face, over the telephone, online, or was self-administered, so the extent of sensitivity bias may depend on this important survey design feature (Hochstim 1967; Tourangeau and Yan 2007). However, we found in our data that survey mode is highly correlated with substantive domain, frustrating our ability to empirically evaluate this theoretical expectation (see Table H8 in the supplementary materials).

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 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 4 and summarize them in Table 4.

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 4 (a), which we walk through in detail for this first example.[14] The top panel of the subfigure presents the estimated level of sensitivity

---

[14]Full-size versions of the panels in Figure 4 with additional study information are included in the supplementary materials.
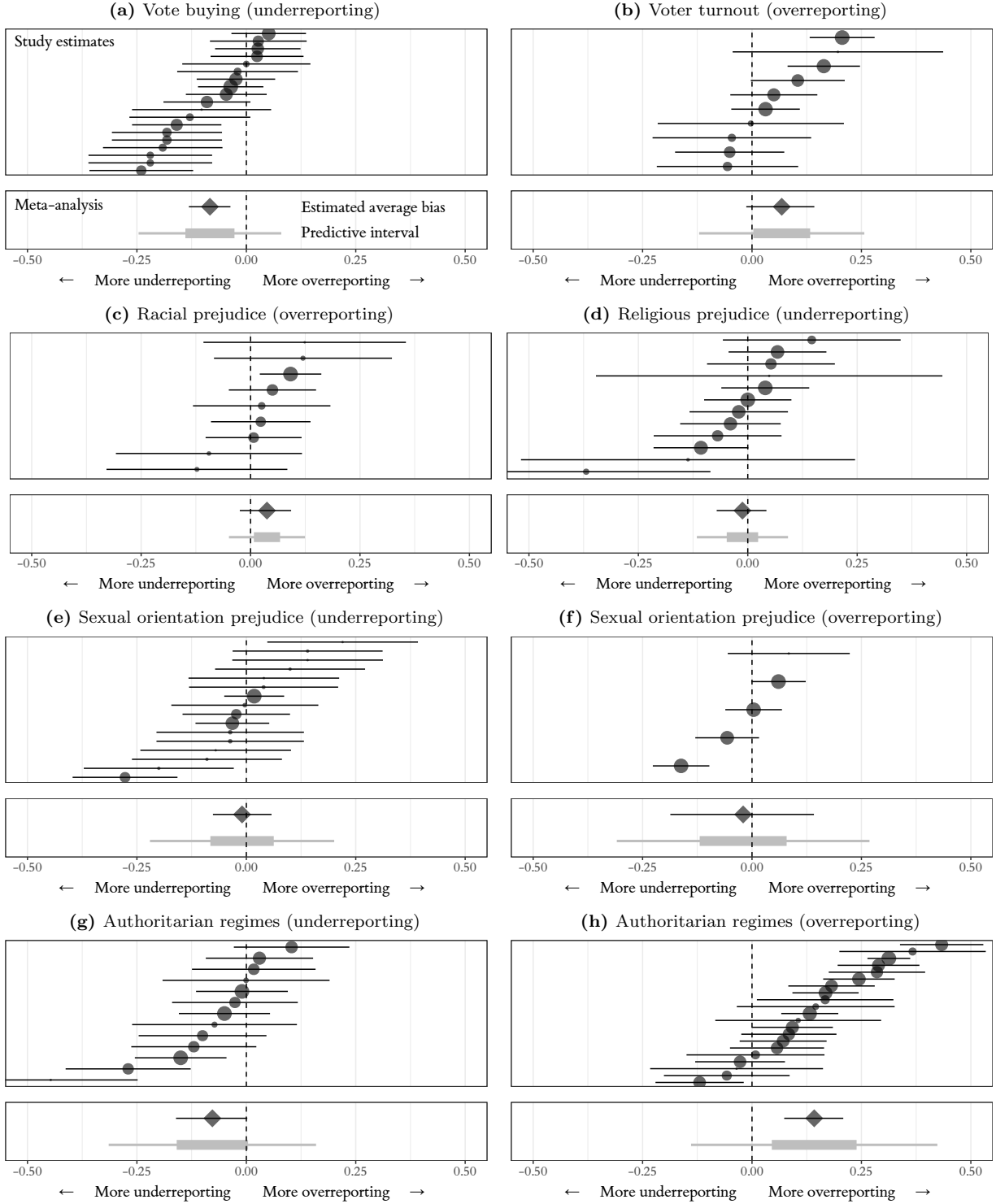
Figure 4: Sensitivity Bias in Four Political Science Research Literatures. Estimated sensitivity bias in each study with 95% confidence intervals, with point size proportional to study weight in meta-analysis (top panel). Estimated average sensitivity bias in each literature (diamond) with 95% credible interval (bottom panel) and estimated 50% (thick gray line) and 95% (thin gray line) predictive intervals for sensitivity bias.

		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 (0.02)	[-0.13, -0.04]	[-0.14, -0.03]	[-0.25, 0.08]	19
Turnout	Overreporting	0.07 (0.04)	[-0.01, 0.14]	[0.00, 0.13]	[-0.12, 0.26]	10
Racial prejudice	Underreporting	0.04 (0.03)	[-0.02, 0.09]	[0.01, 0.07]	[-0.05, 0.12]	9
Religious prejudice	Underreporting	-0.01 (0.03)	[-0.07, 0.04]	[-0.05, 0.02]	[-0.12, 0.09]	12
Sexual orientation prejudice	Underreporting	-0.01 (0.03)	[-0.08, 0.06]	[-0.08, 0.06]	[-0.22, 0.20]	16
	Overreporting	-0.02 (0.09)	[-0.19, 0.14]	[-0.12, 0.08]	[-0.31, 0.27]	5
Support for authoritarian regimes	Underreporting	-0.08 (0.04)	[-0.16, 0.00]	[-0.16, 0.00]	[-0.31, 0.16]	13
	Overreporting	0.14 (0.03)	[0.07, 0.21]	[0.05, 0.24]	[-0.14, 0.42]	21
All results	Underreporting	-0.04 (0.01)	[-0.05, -0.02]	[-0.11, 0.04]	[-0.26, 0.19]	196
	Overreporting	0.12 (0.02)	[0.08, 0.15]	[0.03, 0.20]	[-0.13, 0.36]	68

Table 4: **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.

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, revealing underreporting due to sensitivity bias. For example, in the top row of Figure 4 (a), sensitivity bias is estimated to be 5 percentage points (95% CI -3 to +14) based on a 2010 survey conducted in Bolivia in which the sensitive question read "They gave 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.[15] 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 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 the level of sensitivity bias that would be found in future studies absent additional contextual information. As is clear from the dispersion of the study-level

---

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

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 4 (a), we show the 50% predictive interval from -14 to -2 points and the 95% predictive interval from -27 to +10 points. 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, our posterior estimate based on 10 studies is that voter turnout is overreported by +7 percentage points, but the confidence interval on the meta-analytic average crosses zero. 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 most a small proportion 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, as opposed to memory or recall failures.

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.[16] 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.

---

[16]The focus on these three categories instead of other, equally important subdivisions such as gender or class is entirely due to data availability. 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).

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 nine 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, but the 95% predictive interval admits large negative biases (-16 points) to large positive biases 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 actually do hold their reported views (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-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. 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 an unexpected sign relative to the predictions 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 of +43 points when overreporting is predicted (Kalinin 2015) and a minimum of -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 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 and 0 points. Support for authoritarian regimes is an area where our data suggest there is considerable risk of sensitivity bias. These results are consonant with cross-country comparisons of survey nonresponse rates (Shen and Truex Forthcoming). 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.

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 sufficiently large. We return to two of the main goals of list experimentation: achieving a better RMSE or demonstrating the existence of sensitivity bias.

Figure 5 shows that many list experiments are simply too small. Below the lower curve (98 of 264 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 (107 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
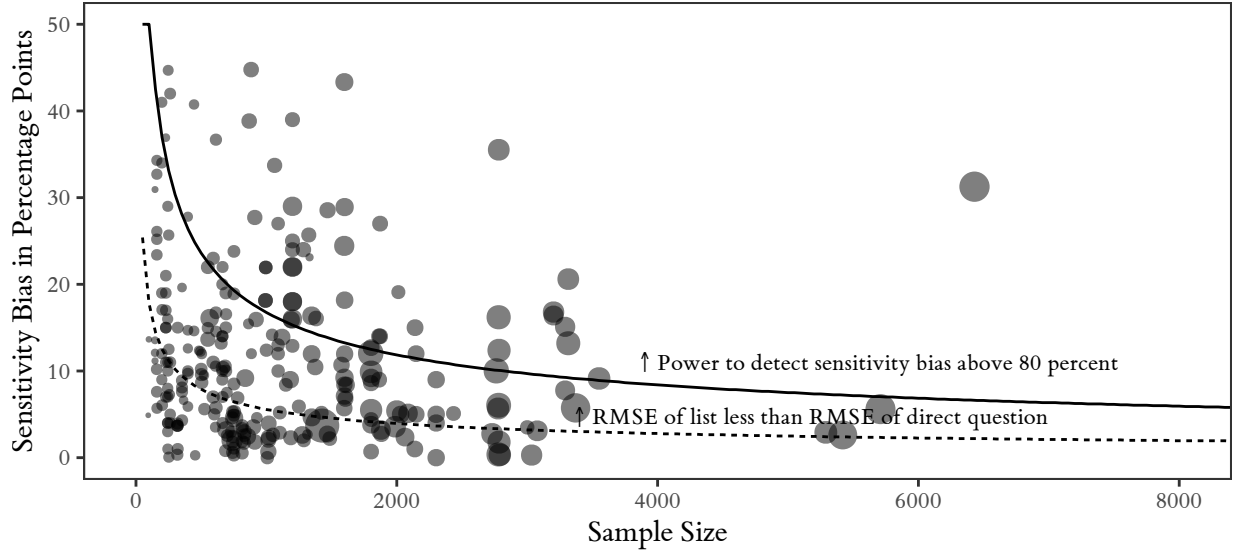
Figure 5: Many Studies of Sensitive Topics Are Smaller than Sample Sizes Recommended based on Power or Root Mean-Squared Error Tradeoffs. Existing studies are overlaid on two design recommendations: the power indifference curve for designs estimating the amount of sensitivity bias, and the RMSE indifference curve for designs estimating the prevalence rate.

direct question, but are not large enough to reliably demonstrate the existence of sensitivity bias. The studies that are above both curves (59 studies) are large enough such that the list experiment is preferred for either purpose.

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.

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 suggestive evidence of overreporting of voter turnout. Surprisingly to us, subjects appear to honestly report their prejudices 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 on survey estimates of prevalence in the first place. The interpretation of the meta-analytic results depends 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.[17] This quantity is still important to learn because applied 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 of 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 as large as 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 wide, which indicates 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

---

[17]For example, no liars may be violated if subjects are simply inattentive (Ahlquist 2018; Alvarez et al. 2019) and no design effects may be violated if control counts depend on the presence or absence of the sensitive item (Flavin and Keane 2009).

in at least some contexts, the technology appears to perform well. In the 166 list experiments for which we have sufficient information, 160 pass the design effects test described in Blair and Imai (2012). 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 because of publication bias or authors' reluctance to share estimates from procedures whose identifying assumptions are not met. If there is a large body of these failed studies, our conclusions about the performance of the list experiment in general might differ.

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 consider four criteria 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 infer their answers exactly or approximately? (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, caution, and further research are of course still warranted.

With respect to what researchers should do about sensitivity bias, the choice between list experiments and direct questions largely amounts to a bias-variance tradeoff. Under typical conditions, list experiments are approximately 14 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. That said, ensuring the safety of participants and research staff may be well worth the loss of precision that comes from list experiments.

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 combines responses with random noise in order to change respondent beliefs about which social referents can infer their responses exactly or approximately. This set includes the list experiment, the randomized response technique, the unrelated question technique, and the crosswise technique. These methods 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 concerned about, these methods will reduce sensitivity bias.

The second set of techniques changes the social 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; Beber et al. 2014), and pre-recorded questionnaires played through headphones 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 participants with their responses, though researchers and state authorities can often later reconnect responses to identifying information.[18] When interviewers or bystanders are the primary social referents respondents worry about, these methods will reduce sensitivity bias as well.

---

[18]For an exception, see Scacco (2012), in which sensitive responses were linked to individuals only through a randomized code that could be reconnected at the researcher's home institution. This procedure was designed to prevent state authorities from connecting responses about participation in riots.

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.

In this paper, we have focused on survey measurement of behaviors as well as explicit attitudes and beliefs. Techniques like the implicit association test (Greenwald et al. 1998) and the affect misattribution procedure (Payne et al. 2005), by contrast, target implicit attitudes. These measures are plausibly free from sensitivity bias, since responses are difficult to consciously misrepresent even if respondents would prefer to hide their views (Nosek 2007, though see Skorinko and Sinclair 2018). Nevertheless, a foundational assumption in the implicit attitudes literature is that implicit and explicit attitudes are fundamentally different constructs. By switching from direct measures of explicit attitudes to indirect measures of implicit attitudes, we may have purged our estimates of sensitivity bias, but we have also changed the research question. This switch is of course appropriate when implicit attitudes are indeed the theoretical target of interest.

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 before turning to the list experiment or similar techniques, mainly because they are so variable that achieving sufficient 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

Adida, Claire, Jessica Gottlieb, Eric Kramon and Gwyneth McClendon. 2019. "Response Bias in Survey Measures of Voter Behavior: Implications for Measurement and Inference." *Journal of Experimental Political Science* 6(3):192–198.

Ahlquist, John S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26(1):34–53.

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27(2):145–162.

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.

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.

Beber, Bernd, Philip Roessler and Alexandra Scacco. 2014. "Intergroup Violence and Political Attitudes: Evidence from a Dividing Sudan." *The Journal of Politics* 76(3):649–665.

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.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113(3):838–859.

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, Kosuke Imai and Yang-Yang Zhou. 2015. "Design and analysis of the randomized response technique." *Journal of the American Statistical Association* 110(511):1304–1319.

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.

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.

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

Coppock, Alexander. 2017. "Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment." *Statistics, Politics and Policy* 8(1):29–40.

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

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.

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

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.

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.

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.

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.

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

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

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.

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.

Hochstim, Joseph R. 1967. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association* 62(319):976–989.

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

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.

Kane, James G., Stephen C. Craig and Kenneth D. Wald. 2004. "Religion and Presidential Politics in Florida: A List Experiment." *Social Science Quarterly* 85(2):281–293.

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.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence and Barbara Mellers. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2):402–419.

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.

Li, Jiayuan and Wim Van den Noortgate. Forthcoming. "A Meta-analysis of the Relative Effectiveness of the Item Count Technique Compared to Direct Questioning." *Sociological Methods & Research* .

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.

Lyall, Jason, Graeme Blair and Kosuke Imai. 2013. "Explaining support for combatants during wartime: A survey experiment in Afghanistan." *American Political Science Review* 107(4):679–705.

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. 2019. *Conditionality and Coercion: Electoral clientelism in Eastern Europe.* Oxford University Press.

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.

Powell, Richard J. 2013. "Social Desirability Bias in Polling on Same-Sex Marriage Ballot Measures." *American Politics Research* 41(6):1052–1070.

Raghavarao, D. and W. T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society: Series B (Methodological)* 41(1):40–45.

Rosenfeld, Bryn, Kosuke Imai and Jacob N Shapiro. 2016. "An empirical validation study of popular survey methodologies for sensitive questions." *American Journal of Political Science* 60(3):783–802.

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

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.

Shen, Xiaoxiao and Rory Truex. Forthcoming. "In Search of Self-Censorship." *British Journal of Political Science* .

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.

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

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