When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments[†]

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

First draft: May 1, 2018
This draft: August 18, 2018

Abstract

Direct survey measures of sensitive beliefs, attitudes, and behaviors may generate biased prevalence estimates. "Social desirability bias" is often invoked as a catchall term to describe the various sources of measurement error associated with sensitive questions. We synthesize work in social psychology and political science on impression management and social desirability to develop a reference group theory of sensitivity bias encompassing both nonresponse and misreporting. We conduct a census of the published and unpublished list experiments conducted to date and compare the results with direct questions. Relative to list experimental estimates, we find that sensitivity biases are typically smaller than 10 percentage points and in some domains, approximately zero. We find that list experiments appear to deliver on their promise of approximately unbiased prevalence estimates. However, in some cases we find that they are unnecessary and are often conducted with samples that are too small. We conclude with specific recommendations for researchers choosing among measurement strategies when asking sensitive questions.

[†]We acknowledge the ISPS Dahl Scholarship, which supported Margaret Moor's work on this project. We thank Jamie Druckman and Larry Hedges, as well as audiences at the Q-Center and the Druckman Lab at Northwestern University for helpful comments.

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

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

[¶]Margaret Moor is a Senior in the College of Yale University.

1. Introduction

Self reports are widely used by social scientists to study quantities that are difficult or impossible to measure without asking a question on a survey. In a small number of well-documented cases, validation studies show that survey estimates can be highly biased. For example, the voter turnout rate in the U.S. is below 60%, but survey estimates put it at between 75 and 85% (Ansolabehere and Hersh 2012). In the other direction, one meta-analysis found that 30-70% of clinically-confirmed recent drug users reported they had not used drugs (Tourangeau and Yan 2007).

Sensitivity bias is the special form of measurement error that stems from the sensitivity of the attitude or behavior being measured; it is theoretically and practically distinct from other sources of measurement error like anchoring, question order effects, or random noise. What makes a topic sensitive depends on the beliefs a subject holds about the relevant reference groups. Sensitivity bias can manifest as nonresponse or misreporting. Respondents may refuse to respond to some kinds of questions in some circumstances, and others may respond but with false, distorted, or imprecise answers.

Determining whether sensitivity bias is a problem in a particular outcome domain is often a matter of intuition, conjecture, or previous theoretical expectations. In a limited number of areas such as turnout and drug use, high-quality validation studies guide assessments of the risk of bias. In other areas, two common heuristics are used. The taboo heuristic considers whether the topic of the question may be uncomfortable for respondents. The social desirability bias heuristic considers asks whether social pressure exists to answer the item in a certain way.

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

Researchers have developed a wide variety of techniques for mitigating the biases associated with sensitive questions. Since the 1950s, when scholars in statistics first identified this problem, cottage industries emerged in nearly every social science discipline to address it. Techniques fall into three broad categories: changing the form of the question (Haire 1950; Warner 1965; Miller 1984), changing the context of how the question is answered (Silver et al.

1986), and measuring which types of people are most prone to giving false answers (Snyder 1987; Paulhus 1991; Berinsky 2004). Despite substantial progress in identifying solutions, there is little to guide researchers in deciding whether to adopt an alternative design and which one to select. Each method comes with costs, in terms of development and testing, survey duration, and statistical power. We develop theoretical guidance comparing a direct question strategy to the most common experimental solution in political science, including the list experiment.

When should we worry about sensitivity bias? In order to answer this question, we define sensitivity bias and consider the bias-variance tradeoff associated with choosing two different formats: asking directly, and asking in a list experiment, a questioning technique designed to minimize misreporting by aggregating responses with several unrelated control items (Miller 1984). Direct questions may be biased but they are low variance; indirect techniques are possibly unbiased, but are higher variance. We compare the theoretical tradeoff with the empirical distribution of sensitivity bias estimates. In short, we find that only when the bias of the direct question is substantial – for a 1000-respondent survey, a bias of 20 percentage points – is the list experiment preferred to a direct question as an estimator of the prevalence rate of the characteristic. The burden is higher still when the researcher wishes to demonstrate that there is misreporting bias by estimating the rate of bias through a comparison to the direct question.

In the meta-analysis portion of our paper, we reanalyze the results of 30 years of experiments that ask questions researchers worry are subject to sensitivity bias. Our research design compares responses to the same question asked directly and as a list experiment.

Our results indicate that sensitivity bias is small for many questions, *contra* the evident expectation on either the authors' or their real or imagined reviewers' parts that misreporting was a concern. However, there is considerably heterogeneity in sensitivity bias across subject domains. Because list experimental estimates are subject to a large amount of sampling variability, we use meta-analysis to aggregate sensitivity bias estimates from many studies. We find evidence of overreporting voter turnout, underreporting vote buying, theft, and racial bias, and nearly no evidence of sensitivity bias in LGBT attitudes. While sensitivity bias can certainly plague direct question estimates, these biases are typically too small to be detected by list experiments with fewer than 3,000 subjects.

2. Theories of sensitive survey responses

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

2.1 Defining sensitivity bias

Scholarly attention to bias from asking sensitive questions directly is motivated by the injunction by early survey researchers not to focus thinking about bias in survey research narrowly on sampling variability but on "total survey error" (Deming 1944). We focus on two types of survey error that may result from asking sensitive questions directly: nonresponse and misreporting, each of which contributes to sensitivity bias.

Misreporting is a type of measurement error in which respondents provide false, distorted, or imprecise answers. Misreporting may represent intentional deception, or it may be due to self-deception, failure to reflect deeply on the true answer, or satisficing (Krosnick et al. 1996; Tourangeau and Yan 2007). The polarity, or direction, of bias is often assumed to be toward a socially desirable or legal response. The polarity of the sensitivity bias need not be constant across respondents.

Nonresponse comes in three forms. Respondents may not participate in a survey if they know its contents are sensitive (unit nonresponse). Gatekeepers or community leaders may also prevent or discourage participation, leading to mass nonresponse (Blair et al. 2014). More commonly, when presented with a sensitive question during a survey, respondents may not answer it (item nonresponse), often recorded as a "refused" or a "don't know" response.

Misreporting and nonresponse due to asking sensitive questions directly is distinct from the generic measurement error and nonresponse that plague all survey research. Here, error and nonresponse are directly related to the latent outcome of interest. In the worst case, surveys on sensitive topics would merely record what the set of subjects willing to respond believe is the socially-expected answer.

2.2 A reference group theory of sensitivity bias

The dominant answer to the question of why people misreport or fail to answer sensitive survey questions identified by social scientists since the 1950s has been social desirability bias (Maccoby and Maccoby 1954). According to Fisher (1993, pp. 303), social desirability

bias results from "the desire of respondents to avoid embarrassment and project a favorable image to others." Goffman's *The Presentation of the Self in Everyday Life* (1959) launched research inquiries across sociology and social psychology into the importance of impression management or self-presentation.[1] Goffman argues that people have in their own minds an idea of how they are perceived by others and take actions to improve that perception. Social desirability bias is a behavioral manifestation of self-presentation. Beyond social desirability, scholars have identified self-image, the fear of disclosure of responses, and intrusive topics as additional causes of sensitivity bias.

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

The last element highlights the fact that we must know *with respect to whom* respondents manage impressions. Psychologists and political scientists have developed and applied scales to measure person-constant levels of desirability bias (Snyder 1987; Paulhus 1991; Berinsky 2004). Yet clearly the set of actors who the respondent believes can hear or read responses, during and after the interview, as well as other elements of the survey context may influence the strength of impression management pressures and the risk of disclosure.

Our theory will relax the assumption that there exists an individual difference variable that distinguishes people on the basis of their *general* susceptibility to engage in misreporting or nonresponse. Instead, we will reinterpret the evidence showing that different people do indeed withhold at different rates in specific scenarios as features of individuals' idiosyncratic beliefs about and perceptions of the relevant reference group. Respondents hold beliefs about *who* is asking questions, who sent the enumerators 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.

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

[1] For a review, see Leary and Kowalski (1990).

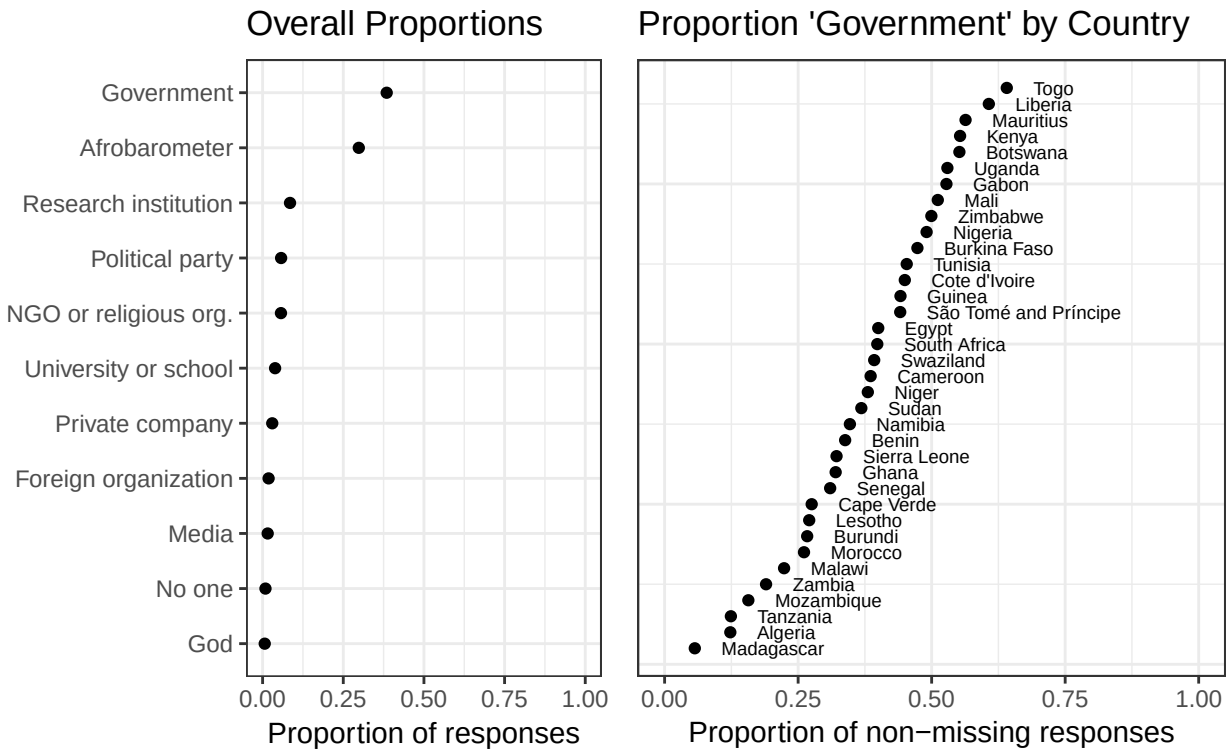

Figure 1: Proportion of Respondents Identifying the Organization Responsible for Afrobarometer Surveys Overall (left panel) and the Proportion Saying the "Government" is Responsible by Country (right panel). The question text is "`Just one more question: Who do you think sent us to do this interview?`" and responses were coded by Afrobarometer from recorded verbatim responses.

countries (right panel). The figure shows that responses vary substantially across respondents, and across countries. Impression management concerns and the perceived risks of disclosure are likely heterogeneous across respondents and context-specific.

Perhaps the most salient reference group for subjects is the enumerator asking the survey question (Feldman et al. 1951). Subjects may presuppose (rightly or wrongly) that survey-takers have an opinion about what the correct attitude to hold is. As a result, trainings for enumerators often include specific guidance for maintaining a neutral demeanor when interviewing respondents (e.g., Glennerster and Takavarasha 2013). Enumerator effects, assessed by randomly assigning subjects to enumerators of a particular type, have been demonstrated for enumerator 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 may con-

stitute a different reference group. Subjects might feel constrained to respond in a particular manner or not at all if under the watchful eye of a spouse, which is the motivation for the frequently-offered survey design advice that surveys be conducted in private if at all possible (Silver et al. 1986; Aquilino 1993; Hartmann 1994; Pollner and Adams 1997; Aquilino et al. 2000; Zipp and Toth 2002). In this case, respondents are less concerned about what the enumerator thinks the right answers are and more worried about what their immediate community think.

Other more distal reference groups may include those who will read responses after the survey ends, such as the sponsoring institution or academic analysts, consumers of the survey data including the media and the general public, or more worryingly, the government or armed groups who might take punitive action depending on the response. Experimental evidence suggests confirms that changing the survey sponsor can change responses to sensitive questions (Corstange 2014).

Existing theory largely focuses on presentation of the self. However, people are not only concerned with how they are perceived by others, but how their *community* is perceived by other communities (Tajfel and Turner 1979). We label this form of sensitivity bias "sociotropic misreporting." For example, individuals may not want the media or the general public to know that their community supports the Taliban (Blair et al. 2014) or their community is afflicted by a high incidence of HIV. Sociotropic misreporting may also be manifested as a respondent belief that if some reference group knew their truthful responses, costs would be imposed on *others*.

Social desirability is not the only source of sensitivity bias. First, respondents face pressures to respond that come from themselves, not only others (Greenwald and Breckler 1985). Second, questions may be seen as "intrusive," representing taboo topics responds feel are out-of-bounds independent of perceived social desirability (Tourangeau et al. 2000). Here, nonresponse may be more common than misreporting. Third, respondents may fear their responses will be disclosed to authorities such as governments, criminals, armed groups, or employers.

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

1. A reference group, or set of people or organizations the respondent has in mind when considering how to respond to a survey question. A reference group could be the respondent him or herself.

2. A respondent perception that the reference group can learn the subject's response to

the sensitive question.

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

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

Using these four criteria, we distinguish sensitivity bias from other forms of measurement error. For example, we draw a fine distinction between self-deception and recall failures. If respondents misreport because they do not want to admit, even to themselves, that they participate in the sensitive behavior, direct questions will suffer from sensitivity bias. If however, respondents simply do not spend sufficient cognitive energy to recall whether, for example, they voted in the most recent midterm election, direct questions will be biased, but not because of sensitivity.

A voluminous research literature examines the distinction between implicit and explicit attitudes (Greenwald and Banaji 1995; Greenwald et al. 1998). Implicit attitudes are unknown even to the respondents themselves, so subjects cannot self-report them. We only consider sensitivity bias that may plague measures of explicit attitudes (Littman 2015).

At a theoretical level, we draw no distinctions here between attitudes and behaviors. Measures of both could be distorted by sensitivity bias, depending on the reference group, the perceived risk of disclosure to that reference group, perceived social norms, and the perceived costs. As an empirical matter, it is unknown whether measures attitudes or behaviors face a greater threat of sensitivity bias.

In the next section, we turn to how a specific measurement tool, the list experiment, may or may not alleviate sensitivity bias by making a respondent's survey responses invisible to a particular reference group.

2.3 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 it with the answers to several binary control items. Sensitive item responses, thus, are hidden from enumerators, bystanders, and data consumers who only hear the count including control items. Researchers can estimate the prevalence rate of the sensitive item by subtracting the average count of the control items from the observed count, leaving only the sensitive item.

Why does this address misreporting and nonresponse due to the sensitivity of the question? The list experiment is designed to minimize self-presentation concerns and the risk of disclosure. Self-presentation concerns of the respondent are addressed by avoiding revealing the sensitive item response to the enumerator, bystanders, or later data consumers. Presentation concerns about self-image are not addressed, however. The list experiment also minimizes the risk of disclosure. Authorities, such as employers or parents, cannot exactly identify the sensitive item response for most respondents. (We discuss the cases in which responses are exactly identified in a moment.) The list experiment does not change the bias from asking intrusive questions, because the text of the question includes the same sensitive item text found in the direct question.

The list experiment obscures individual responses to the sensitive item, but still allows analysts to estimate sample quantities including sensitive item prevalence and other relevant quantities. To illustrate how, we introduce a minimum of formalism following Blair and Imai (2012). With a set of N individuals indexed by i, we randomly assign each to a treatment group identified by $T_i = 1$ or a control group identified by $T_i = 0$. In the control group, we ask respondents for a count of the number of "yes" responses to J control items indexed by j. In the treatment group, we ask respondents for a count of the number of "yes" responses to a set of $J + 1$ items, the J control items plus the sensitive item. We define two sets of potential outcomes. We represent the "yes" or "no" responses to each item as $Z_{ij}(t)$ for $t = 0, 1$, i.e. the latent response to item j for respondent i in treatment group t. We then represent the observed counts of "yes" responses as $Y_i(t)$. The control potential outcome $Y_i(0) = \sum_{j=1}^{J} Z_{ij}(0)$ ranges from 0 to J and the treatment potential outcome $Y_i(1) = \sum_{j=1}^{J+1} Z_{ij}(1)$ ranges from 0 to $J+1$ given the addition of the sensitive item. The observed outcome is defined as $Y_i = Y_i(T_i)$.

A main aim of researchers is to estimate the sample prevalence of the sensitive item, i.e. $\tau \equiv \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$. In order to identify this quantity, four assumptions must be invoked. These are described in Imai (2011), but we recapitulate them here. First, the standard assumptions for identifying the average treatment effect in an experiment: noninterference (sometimes referred to as SUTVA, or the Stable Unit Treatment Value Assumption) and the ignorability of the treatment status. Both are typically guaranteed by design.[2] These two

[2]Noninterference requires that subjects' outcomes depend only on their own treatment status and not on that of other subjects. In list experiments (and in survey experiments in general), noninterference is typically assured by design because subjects take the surveys separately. Ignorability requires that the treatment be independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$ and is guaranteed by design in list experiments because the treatment is randomized.

assumptions are required to identify the average treatment effect in observed counts.

Two additional assumptions are required in order to interpret this treatment effect as the sensitive item prevalence rate. Both assumptions are exclusion restrictions. No design effects assumes that responses to the control items do not differ in treatment and control. No liars assumes that respondents do not misreport the "yes" or "no" response to the sensitive item within the count.[3] Substantively, no liars means that list experiment responses are distorted by sensitivity bias; the protection provided by the list experiment removes the threat of costs because the reference group cannot learn subjects' responses. No liars would be violated if subjects were still unable to admit the truth to themselves.

Under noninterference, ignorability, no design effects, and no liars, the sample sensitive item prevalence is nonparametrically identified. We estimate this quantity using the difference-in-means estimator, which is an unbiased estimator under these assumptions. Robust HC2 standard errors are justified by the design.[4]

For political scientists, quantities of interest beyond the sensitive item prevalence have also been of interest. Subgroup prevalence (analogous to conditional average treatment effects in standard experimental settings) and their differences can be estimated with the same tools and justifications. For surveys that also include a direct question on the same topic, like the Kenya postelection survey reported in Kramon (2016), the difference between the list experiment estimate and the direct question estimate of the sensitive item propensity is often of interest. The difference is sometimes labeled an estimate of social desirability bias (Janus 2010; Blair and Imai 2012), though it is better thought of as an estimator of sensitivity bias more generally.

2.3.1 Example: Vote Buying in Kenya

We illustrate the design with an example. Kramon (2016) reports on a post-election survey after the 2007 Kenyan election of a nationally-representative sample of $2,000$ Kenyans. The survey focuses on estimating the proportion of voters who experienced vote buying during the election. To do so, the authors use a list experiment and a direct question. Respondents were randomized into two groups.[5] In the control group, respondents were read the following

[3]Formally, the no design effects assumption states that for all respondents i, $\sum_{j=1}^{J} Z_{ij}(0) = \sum_{j=1}^{J} Z_{ij}(1)$. The no liars assumption states that for all respondents i, $Z_{i,J+1}(1) = Z^*_{i,J+1}$.

[4]Other estimators, including ordinary least squares with covariate adjustment or the nonlinear least squares estimator proposed in Imai (2011), invoke additional modeling assumptions in order to generate more precise estimates and to enable to the estimation of other quantities of interest including multiple regression coefficients.

[5]A third group saw a treatment list with a different sensitive item. This group is included in the results section below but is ignored for the purposes of this example.

instructions:

> Election campaigns are a busy time in our country. I am going to read you a list of some of things that people have told us happened to them during the 2007 campaign. I am going to read you the whole list, and then I want you to tell me how many of the different things happened to you. Please do not tell me which of the things happened to you, just how many. If you would like me to repeat the list, I will do so.
>
> 1. Politicians put up posters or signs in the area where you live.
> 2. You read the newspaper almost every day to learn about the campaign.
> 3. You met a politician personally to discuss his or her candidacy.
> 4. You discussed the campaign with a chief or another traditional leader.

Responses in this control group represent the total number of control items the respondent answers "yes" to. In the "Influenced" treatment group, the same script was read but with a fifth item added to the list:[6]

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

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

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

Thus, in the treatment groups, responses represent the count of the number of "yes" responses to the set of control items and the sensitive item. The observed data for treatment and control are displayed in Table 1.

Using data from the Kramon (2016) postelection survey in Kenya, we can estimate the prevalence rate of vote buying, the main quantity of interest in the study. The author also investigates whether vote buying is more common in localities where political parties can monitor individual votes, an implication of classical theories of vote buying. We estimate the conditional prevalence rate between areas with and without the ability to monitor, and its difference.

Figure 2 presents our results. The "influence" question appears to be affected by sensitivity bias: the list experiment estimate, while imprecisely estimated, is definitively higher

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

Count	Control	Received Treatment	Influenced Treatment
1	290	235	215
2	235	280	204
3	72	96	113
4	25	30	29
5	0	12	8

Table 1: Observed list experiment responses by treatment status for whether a bribe was received (third column) and whether the bribe influenced the respondent's vote (second column) from the 2007 Kenya postelection survey reported in Kramon (2016).

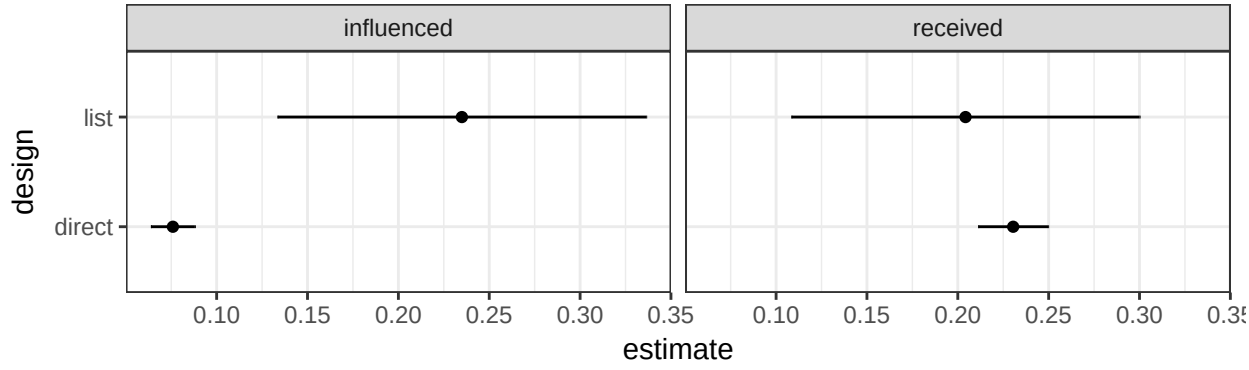

Figure 2: Estimated prevalence of vote buying from the list experiment and the direct question, and conditional prevalence rates among localities according to the strength of party ability to monitor voting.

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.

2.3.2 Violations of the identifying assumptions

In most settings, the random assignment and noninterference assumptions are assured by design, but the same cannot be said for the no liars and no design effects assumptions.

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

Blair et al. (2014) provide direct evidence of both floor and ceiling effects in their study of respondents in rural parts of five war-affected provinces in Afghanistan. Zero out of $2,756$ respondents answered "none" in the treatment group, and zero answered $J + 1$ items. As an aside, these data also indicate that these relatively uneducated respondents were able to understand that the list experiment gives them cover – precisely because they avoided response options that would exactly identify their position on the sensitive item.

Scholars have devised design-based and model-based approaches to address ceiling and floor effects. Glynn (2013) provides advice on selecting control items to minimize the number of respondents who would respond "all" or "none" in the treatment group: choose negatively-correlated items, a high propensity item, and a low propensity item. Blair and Imai (2012) develop maximum likelihood models to adjust for ceiling effects, floor effects, or both. Ahlquist (2018) proposes alternative forms of misreporting due to satisficing, and Blair et al. (Forthcoming) propose maximum likelihood models to address them.

Even in the absence of floor or ceiling effects, no liars might be violated. The higher the count in the treatment group, the higher the probability that the sensitive item response is a "yes." Respondents who perceive that their list response provides partial information about the sensitive item may "lie" in the sense of decrementing their true list response by one.

Respondents may also (correctly or incorrectly) believe that enumerators and bystanders may be able to predict the control item count based on respondent characteristics. When items are negatively correlated and high-and low-prevalence items are included, the accuracy of these predictions increase. For example, if half control items are things most Democrats would agree with and the other half are things most Republicans would agree with, then predicting the control item count based on respondent partisanship is easy.

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

Design effects may also be induced by the presence of the sensitive item in the treatment group list due to its sensitivity. Scholars worry that adding the sensitive item triggers impression management concerns generally, and that may spill over into the control items.

	Bribe Received				Bribe Influenced Vote			
	$Z_i = 0$		$Z_i = 1$		$Z_i = 0$		$Z_i = 1$	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
$Y_i(0) = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$Y_i(0) = 1$	0.11	0.03	0.36	0.02	0.09	0.03	0.38	0.02
$Y_i(0) = 2$	0.06	0.02	0.32	0.03	0.11	0.02	0.27	0.03
$Y_i(0) = 3$	0.02	0.01	0.09	0.02	0.02	0.01	0.09	0.02
$Y_i(0) = 4$	0.02	0.01	0.02	0.01	0.01	0.00	0.03	0.01

Table 2: Estimated proportions of latent responses to the control items and the sensitive item from the two list experiments reported in Kramon (2016). Negative estimated proportions would indicate a violation of the no design effects assumption. No negative estimates obtain.

Zigerell (2011) notes that respondents may want to send a strong signal that they do are not answering the sensitive item in the affirmative by deflating their responses to the control items to be closer to or at a zero response. For this reason, scholars have noted that the control items need not be nonsensitive (Blair and Imai 2012), and indeed it may be preferable if the items are similar in content and sensitivity to the sensitive item.

Like any excludability assumptions, the no liars and no design effects assumptions are not directly testable and their credibility in any given research scenario must be justified with theoretical argument and qualitative evidence. That being said, a statistical test of the null hypothesis of no design effects was proposed in Blair and Imai (2012).[7] The test relies on the fact that, under the four assumptions outlined above, the joint distribution of the potential outcomes $Z_{ij}(0)$ and the sensitive item $Z_{i,J+1}(1)$ is identified. We can estimate the proportion of respondents that fall into each possible cell representing a pair of a control item count (from 0 to J) a sensitive item response (0 or 1). When those counts are estimated to be negative, this indicates that respondents' answers to the control item counts differ across treatment and control – a violation of the no design effects assumption. To illustrate, we present estimates of each cell in Table 2, all of which are estimated to be positive. The p-value of the test described in Blair and Imai (2012) is 1, indicating a failure to reject the null of no design effect.

[7] Aronow et al. (2015) combine direct questions and list experiments to motivate an alternative test of the no liars and no design effects assumptions. Among those who directly *admit* to the sensitive behavior, the list experimental estimate should be equal to one – rejecting the null hypothesis that it is is equivalent to rejecting at least one of the no liars, no design effects, and "no false confessions" assumptions.

3. Tradeoffs in the Choice of a Measurement Design

When the four criteria of sensitivity bias are met and researchers wish to mitigate the risk of bias, the list experiment is a tool that may reduce bias for reasons described in the preceding section. However, the list experiment presents a well-known bias-variance tradeoff. Direct questions may be biased, but they produce low-variance estimates. Under the assumptions described above including no liars and no design effects, list experiments are unbiased, but they produce relatively-high variance estimates. In this section, we characterize theoretically the bias-variance tradeoff between several designs that rely either on direct questions or the list experiment. In the next section, we assess with a meta-analysis where on this bias-variance curve existing studies fall.

Consider a study of $N = 2,000$ subjects with a true prevalence rate (π^*) of 50%, but that has a misreporting bias (b) of the direct question of 10 percentage points. Y_i is the response that subject i gives to the direct question. The direct question estimator $\hat{D}$ is the sample mean, i.e. $\hat{D} = \frac{1}{N}\sum_1^N Y_i$. As we demonstrate in the appendix building on Samii (2012), the variance and standard error $\hat{D}$ are given by,

$$\mathbb{V}(\hat{D}) = \frac{\pi^* * (1 - \pi^*) + b * (1 - b) + 2 * (b - b * \pi^*)}{N - 1}$$

Plugging in the assumed values of $N = 2,000$, $\pi^* = 0.5$, and $b = 0.1$ and taking the square root yields a standard error of 0.015, or 1.5 percentage points.

By contrast, the variance and standard error of the list experiment conducted among the same set of subjects is much larger. As shown in the appendix, these are given by

$$\mathbb{V}(\hat{L}) = \frac{4 * Var(Y_i(0) + \pi^* * (1 - \pi^*)}{N - 1},$$

where $\mathbb{V}(Y_i(0))$ is the variance of the control item response. Assuming the variance is equal to 1 (Glynn 2013, p. 163, footnote 6), we obtain a standard error for the list experiment of 0.046, or 4.6 percentage points. For the same number of subjects, the list experiment is $(4.6/1.5)^2 \approx 10$ times more variable than the direct question. Stated differently, a researcher would need a sample of 20,000 subjects in order to produce a list experiment estimate as precise as the direct question with 2,000. The intuition for this stark shortcoming of the list

experiment is that only half the sample is asked about the sensitive trait and their responses are further obscured by adding noise (the control count). At the end of this section, we discuss innovations in the design of list experiments that aim to mitigate this difficulty.

In what follows, we consider how the properties of these two designs, the direct question and the list experiment, inform the decision to undertake a list experiment. The choice to do a list experiment will depend on the goal of the research. We identify three main goals:

1. To estimate a prevalence rate as well as possible, in terms of root mean squared error (RMSE).

2. To demonstrate the presence of misreporting bias, via a hypothesis test against the null of no difference between the direct and list estimates.

3. To estimate the difference in prevalence rates across experimental or nonexperimental groups.

3.1 Estimation of prevalence rate

In some research settings, the primary goal is obtaining a good estimate of the overall prevalence rate of a sensitive trait, as in Gervais and Najle (2018), which sought to estimate the proportion of the U.S. that is atheist. It is unclear, ex ante, which measurement procedure, the direct question or the list experiment, will render estimates that are closer to the true prevalence rate. The direct question is low variance, but thought to be biased, the list experiment is unbiased, but will have a higher variance. The main parameters that govern which approach will have a lower root mean squared error (RMSE) are the extent of bias and the sample size of the study.

Figure 3 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 misreporting bias. The line in the figure describes the bias/sample size combination at which researchers should be indifferent between the two methods on the basis of RMSE. For a study with 1,000 subjects, the bias must be greater than 6 points to prefer a list experiment; at 2,000, the bias must be greater than 4.5 points.

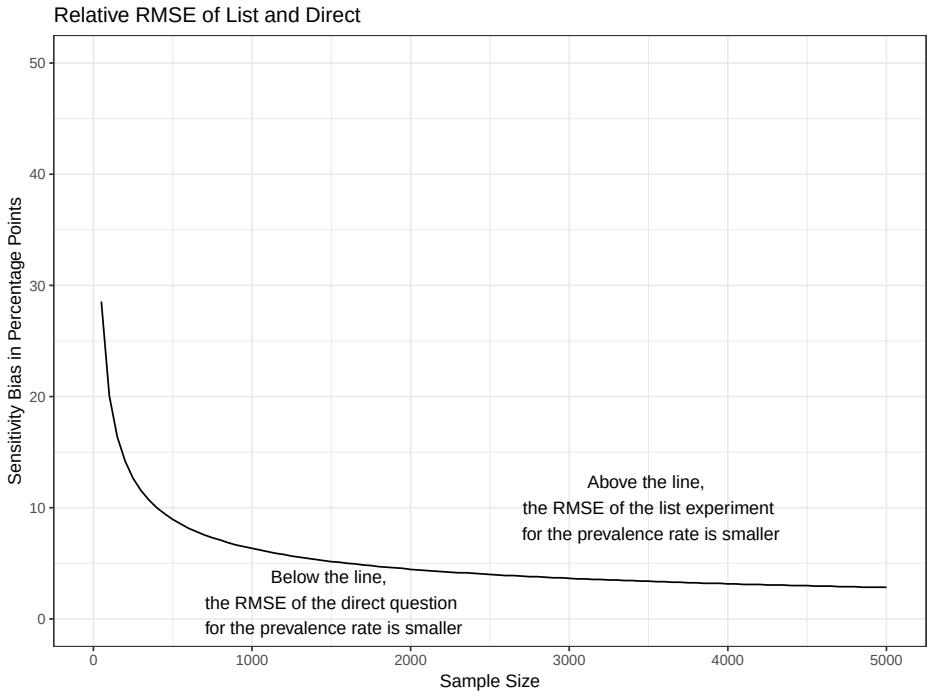

Figure 3: Difference in root mean squared-error between list experiments and direct questions

3.2 Demonstrate existence of misreporting bias

Another goal in some settings is to show that a particular domain is or is not plagued by misreporting bias. Lax et al. (2016) conduct a list experiment and a direct question to estimate support for same-sex marriage; they find no difference and conclude that direct questions produce trustworthy measures of attitudes about same sex marriage. Coppock (2017) uses a similar design to demonstrate the apparent absence of "Shy Trump Voters."

Typically, researchers will first estimate the difference across questioning modes, then conduct a hypothesis test against the null of no difference. The higher the bias or sample size, the higher the power of this design to detect misreporting bias. However, figure 4 shows that the power is generally quite poor. The figure plots the bias / sample size combinations at which the power to detect misreporting bias is 80%. At 1,000 subjects, the bias would need to be 20 percentage points in order to reach 80% power; even at a sample size of 2,000, power to detect biases of 10 percentage points is well below the conventional power target.

3.3 Group differences in prevalence rates

Many social scientific theories predict that prevalence rates will differ according to subgroups defined by individual-level covariates such as race, gender, or political orientation. Further,

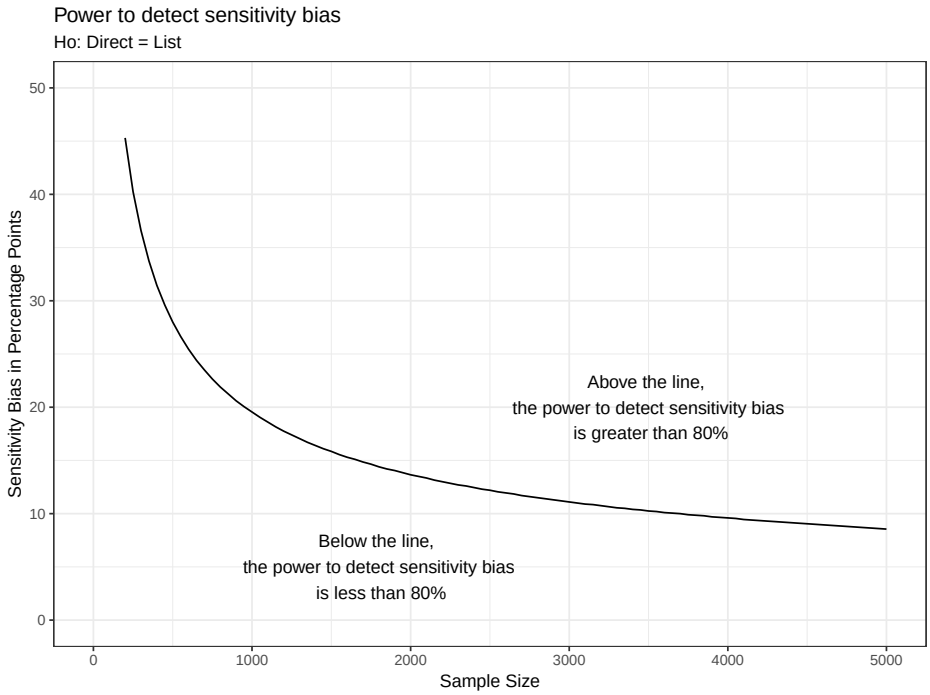

Figure 4: Statistical Power to Detect Withholding Bias

some experimental interventions are designed to *change* whether or not a person holds an attitude or engages in a behavior. In both the observational and experimental cases, a common concern is that the differences in prevalence rates obtained by a comparison of the average direct question response across groups are biased due to misreporting. In such cases, a common practice is to estimate the difference in prevalence rate via a regression of the list experiment response on treatment, and indicator for group membership, and the interaction between the treatment and group membership indicators. The coefficient on the interaction term is an estimate of the difference-in-prevalence rates.

As described in Samii (2012), the high variance of the list experiment frustrates the comparison of prevalence rates across groups, regardless of whether those groups are formed experimentally or on the basis of background attributes. Figure 5 shows that the power to detect even substantial differences in prevalence rates is abysmal. Differences must exceed 25 percentage points before a 2,000 unit study has 80% power to detect them; they must be 20 points or more in the case of a 3,000 unit sample. Conclusively demonstrating that two groups have different prevalence rates requires extreme differences and very large samples.

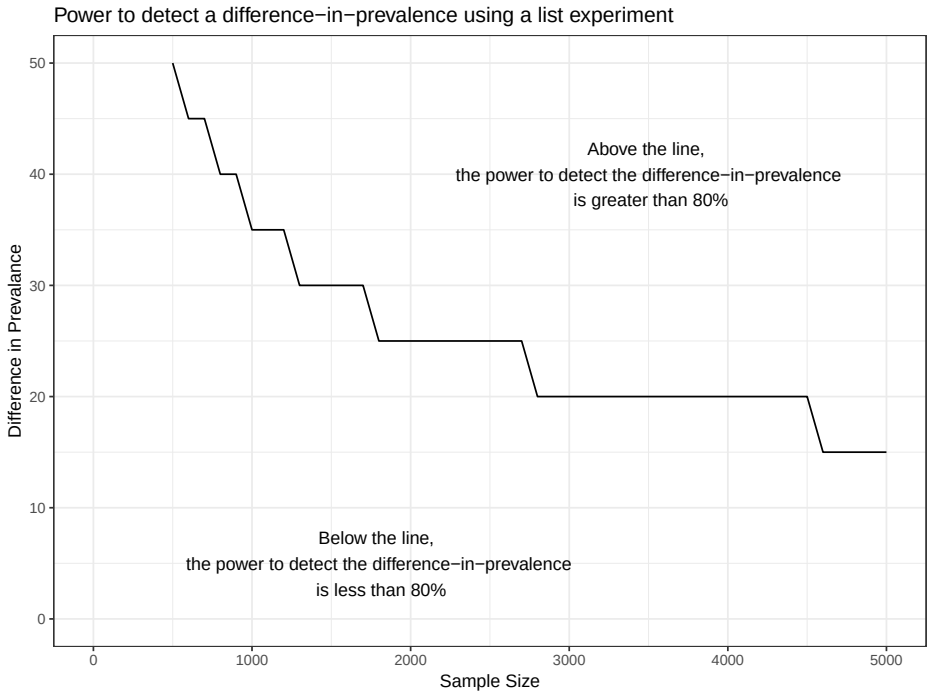

Figure 5: Statistical Power to Detect A Difference-in-Prevalence

3.4 Using list experiment estimates as a predictor

Every quantity of interested presented thus far has the list experiment as the outcome of study. In some applications, the sensitive item is a predictor, not an outcome. A simple analysis would construct predicted probabilities of the list experiment, for example from a linear probability model, and include those predicted as a predictor in the regression of interest. Imai et al. (2014) propose several more efficient estimators that build on this intuition. We leave to future research examining the bias-variance tradeoff in using the list experiment as a predictor compared to using direct question responses.

3.5 Improving power of the list experiment design

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

Droitcour et al. (1991) proposes the double list experiment in which all subjects participate in two list experiments with different control items but the same sensitive item.

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

Other scholars have proposed various methods for combining list experiments with other sources of information. Blair et al. (2014) proposes a combined list and endorsement experiment that succeeded in reducing the variance of the list experiment by 12%. Aronow et al. (2015) derive a method for combining list and direct questions by conducting a list experiment among those subjects who do not directly admit to the sensitive trait. This procedure recovers the precision of the direct question among those whose answers are presumed to be truthful and combines it with the unbiased estimate generated by the list experiment among those who do not admit. In their applications, the combined estimator decreased variance by 12% to 50%. Chou et al. (2018) provide a generalization of Aronow et al. (2015) to any subgroup among whom the true prevalence rate is known. Auxiliary information of this sort, though rare, can dramatically reduce variability. In their application to support for an antiabortion ballot measure, auxiliary information in the form of known vote totals reduced the variance of the list experiment by an estimated 88%.

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

Table 3 shows how each of these methods help to decrease variance. The final column shows what the sample size of the standard list experiment design would need to be in order to achieve the same precision as each method conducted on a sample of 2,000 subjects. The feasibility of each improvement will vary depending on the application; sometimes unavoidable features of the setting will cause violations of the specific assumptions invoked by each

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

Table 3: Variance reduction and increase in effective sample size (relative to a N = 2,000 standard design) of alternative list experiment designs

design. For example, if misreporting bias affects different subgroups in opposite directions, the required assumption of monotonicity in Aronow et al. (2015) would be violated and the method cannot be used. If the sensitive trait is a behavior, it can be difficult to construct an endorsement experiment – in such cases, the proposal in Blair et al. (2014) would not be feasible.

4. Research design

When should researchers worry about sensitivity bias? If it is a risk, is the list experiment the right tool to mitigate the risk? We conduct a meta-analysis of list experiments to provide answers to both questions. We estimate the level of sensitivity bias in domains that span the social sciences, by comparing responses to the same question asked directly and asked as part of the list experiment. We then present the empirical distribution of estimated bias and sample size. We use this evidence to assess where on the bias-variance tradeoff identified in the last section as the way to choose between asking direct questions and using the list experiment.

We attempted a census of all list experiments ever conducted, published or otherwise, as of December 31st, 2017. We certainly failed in this task. At a minimum, we have heard from colleagues of list experiments they ran that were never written up and whose data is long since lost. We searched Google Scholar, SSRN, the Harvard Dataverse, and political science conference programs for the past seven years with the search terms "list experiment," "item count technique," "unmatched count technique," and "double list experiment." We then

published this list of studies on www.sensitivequestions.org and on the authors' personal Twitter accounts with a call for any additional studies that we might have missed. This search yielded 147 papers containing 469 distinct experiments.

We collected general metadata on each study and classified the sensitive question topic into a series of categories modeled on Blair and Imai (2012). We hand-coded the following meta-data about each study: the discipline, substantive category and subcategory, the survey mode (in person, online, telephone, or other), and the polarity of the sensitive item, i.e., the hypothesized direction of the misreporting bias. This process involved stipulating a reference group, the supposed preferences of the group, and the costs a subjects would face for giving the "wrong" answer. While we acknowledge that misreporting bias could, in principle, cut both ways depending on subgroups, we encountered no articles that made strong theoretical or empirical claims to that effect. We categorized each direct question as being prone to overreporting, prone to underreporting, or nonsensitive. We relied on the theoretical arguments in made in the original articles where available, and substituted our best judgment in the remainder of cases.

We gathered statistical information about the list experiments using a cascading data collection approach. In the best case, we obtained the replication dataset from online journal appendices, the Harvard Dataverse, authors' personal websites, or private communication. When replication data were not available, we searched the paper for a crosstab of list experiment responses by treatment condition (similar to 1). These crosstabs contain sufficient information to reconstruct all required statistics. For each list experiment for which we had data or could reconstruct from a crosstab, we calculated the difference-in-means estimate of prevalence and standard error. Finally, if neither the data nor the crosstab was available, we searched the paper for the estimated prevalence rate and standard error. In rare cases, a study reported a prevalence rate estimate but no standard error; in those cases we imputed our best guess for what the standard error would likely be from the sample size. These guesses derive from a regression of the estimated standard error on sample size among studies for which both figures were available. This admittedly crude procedure both comes close to the theoretical standard errors under reasonable assumptions and does a moderately good job of predicting the standard error in those cases where it is known ($R^2 = 0.54$).

The direct question estimates of prevalence all come from the original authors. Some studies obtained direct estimates by asking the direct question to either their entire sample or a random subset; others referred to an estimate obtained by other scholars or agencies. We logged whichever direct estimate was reported by the original authors. We grant that

this procedure is subject to selection bias. If authors preferred to report direct estimates that are either very different from or very similar to the eventual list estimate, then the resulting meta-analytic picture would be distorted. However, we elected not to independently obtain direct prevalence estimates (e.g., from publicly-available surveys) as such discretion could lead to the perception that we were seeking to obtain a pattern either favorable or unfavorable to list experiments.

Publication bias arising from the file-drawer problem (Gerber and Malhotra 2008; Franco et al. 2014) has, anecdotally, been a problem in the list experiment literature. We might reasonably assume that papers featuring list experiments would be easier to publish if the method provides a novel or surprising result. In most circumstances, this means demonstrating that the list estimate is different from the commonly accepted estimate or one derived from direct questions. In the course of our data collection, we heard from many scholars who claimed to have "not found anything" when using a list experiment. We tried hard to overcome the publication filter bias by seeking both published and unpublished experiments.

We also acknowledge a second source of selection that may have influenced our results: the direct question estimates. In all cases, we relied on the original paper to provide direct question estimates. In many cases, direct questions were not available; by no means is this missingness at random. In cases where direct questions were available, we must entertain the possibility that authors (consciously or unconsciously) chose direct question estimates that were favorable or unfavorable to the authors' view of misreporting bias in their survey context. If publication and career incentives at play favor demonstrations of bias over demonstrations of limited bias, we expect both that the "missing" direct questions are those that do not disagree with the list experiment estimate and that the direct question we do have would be selected to be maximally different from the list estimate. We acknowledge that arguments that the bias is understated could also be made in at least some domains.

4.1 Assessing the assumptions of the list experiment

We rely on the list experiment estimate of the prevalence of the sensitive item as a measure of the true prevalence rate. To do so, we invoke the four assumptions described in Section 2.3. Before we conduct the meta-analysis, we assess the validity of these assumptions. In particular, we examine the key no design effects assumption.

For each list experiment we collected, we perform a test of the no design effects assumption from Blair and Imai (2012), described above in section 2.3. Out of 160 list experiments for which we have sufficient information to conduct the test, only 6 fail. The test itself may be

somewhat underpowered to detect small violations of no design effects, but the overwhelming success rate suggests at a minimum that large violations of no design effects are implausible. We include the few studies that fail for fear they are false positives; our conclusions are robust to the inclusion or exclusion of these few studies.

4.2 Meta-analysis of sensitivity bias

With the list experiment and direct question estimates in hand, we estimated the difference as an estimate of misreporting bias. We estimated the standard error of the difference between the direct and list estimate assuming independence of the two estimates.[8] We calculated a 95% confidence interval for the difference under a normal approximation, i.e. we added and subtracted $1.96 \cdot$ SE(difference) from the estimated difference.

In interpreting the difference between list experiments and direct questions as a measure of sensitivity bias, we make several assumptions in addition to the standard assumptions of the list experiment. We assume no differential nonresponse between questions. We assume there are no order effects. We assume that differences in question wording of the sensitive item do not affect responses. Finally, we assume that the list experiment and direct question were asked of the same sample, or of two samples from the same population.

5. Results

We present two sets of results. First, we provide an answer to the question, when should we worry about sensitivity bias? We present meta-analytic estimates of the degree of sensitivity bias across many subject domains in the social sciences. Second, if sensitivity bias is present, we present empirical evidence to identify where on the bias-variance tradeoff between list experiments and direct questions. This empirical distribution helps answer the question, should I use a list experiment to mitigate the risk of sensitivity bias?

5.1 Estimates of Sensitivity Bias

Our database of pairs of list experiments and direct questions is highly heterogeneous, so we will present results in a series of figures grouped by substantive topic area. In the appendix, we provide the direct question, the sensitive item, the two prevalence estimates, and their

[8] Specifically, we calculated SE(difference) $= \sqrt{\text{SE(list)}^2 + \text{SE(direct)}^2}$. This formula assumes that the direct and list estimates are independent; this assumption will be mildly violated if both the direct and list estimates are calculated on the same sample. Under the assumption that direct and list estimates are positively correlated, our naive estimates of sampling variability are conservative.

difference for each study separately. We present here six topic areas most relevant to political science: nonsensitive questions, theft and fraud, racial prejudice, LGBTQI attitudes, vote buying, and turnout.

Each figure displays the direct estimate minus the list experiment estimate, with a 95% confidence interval around the difference. Under the list experiment assumptions, higher values indicate that subjects *overreport* the sensitive behavior or attitude when asked directly and lower values indicate that they *underreport*. We have categorized the direct questions according to whether, according to the original authors and in our own judgement, sensitivity bias should be positive or negative.[9] Interpreting the difference between list and direct as an estimate of sensitivity bias depends on the list experiment assumptions as well as the ancillary assumptions enumerated above.

We begin with questions that the original authors deemed "nonsensitive." In terms of our four necessary conditions for sensitivity bias, we think these questions are nonsensitive either because the reference group (presumably the survey enumerator or the researchers) do not have a preference over the answers or there are obviously no costs associated with giving a dispreferred answer, if it exists. This category of questions includes topics like watching CNN or brushing one's teeth. The comparison of list and direct estimates of prevalence presented in Figure 6 confirms these intuitions: on average, the difference between the two question formats is close to zero. The fact that the answers do not appear to diverge provides a validation that at least in these experiments, the no liars and no design effects assumptions do not appear to be violated.

By contrast, direct question measures of the prevalence of theft and fraud appear to be badly biased downward by sensitivity bias. Figure 7 shows that the list estimate is usually (but not always) higher than the direct estimate in this topic area. On average, the difference is 9 percentage points (SE: 2.3 points). When subjects face the clear threat of a legal consequence (in addition to the social disapproval from the enumerator that they may be avoiding), they withhold true answers. Further, the list experiment appears to provide enough cover for subjects to be honest.

One of the earliest successes of the list experiment was the study of racial prejudice, beginning with Kuklinski et al. (1997). Figure 9 shows a curious pattern of results. We do find suggestive evidence of overreporting support for black politicians or a family member marrying a black person. We do not, however, find evidence that respondents underreport

[9]We actively investigated the possibility that sensitivity bias might cut in different directions for different subjects. In our dataset of papers, we found no instances of original authors claiming cross-cutting biases, either theoretically or empirically.

racist attitudes. In fact, on average, the list experiment estimates are *below* the direct measures, running counter to the notion that subjects are unwilling to directly report racist views. Figure 8 shows that the pattern observed for racial attitudes does not appear to extend to LGBTQI attitudes. Neither attitudes that ought to be overreported nor those that ought to be underreported exhibit signs of sensitivity bias.

Lastly, we turn to two important political behaviors, turnout (Figure 11 and vote buying (Figure 10). Consistent with theoretical expectations, we find that turnout is overreported by an average of 7 points (SE: 3 points) and that vote buying is underreported by an average of 8 points (SE: 2 points). Meta-analysis is an important tool for increasing precision in these two cases, as many of the study-level estimates of sensitivity bias are very imprecise.

5.2 Empirical Distribution of Estimated Bias and Sample Size

The corpus of list experiments described in the previous section provides an means to gauge the extent to which list experiments conducted to date are of a sufficiently large size for either goal: achieving a better RMSE or demonstrating the existence of misreporting bias.

The graph shows three regions. Below the lower curve, 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, the choice between list experiments and the direct question depends on the goal of the research. These list experiments are large enough to produce lower RMSE than the direct question, but are not large enough to reliably demonstrate the existence of misreporting bias. The studies that are above both curves are large enough to be preferable for either purpose. Figure 12 shows that many list experiments are simply too small.

We emphasize, as discussed in Section 3.5, that the indifference curves between list experiments and direct questions included in Figure 12 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. The effective sample of each study can be estimated using the deflators in Table 3 for each design and analysis modification.

6. Discussion

Scientific survey research traces its origins to George Gallup's first nationwide sample survey in 1936. Shortly thereafter, scholars began to worry if the answers they obtained suffered from misreporting and nonresponse due to the sensitivity of some questions.

Our main goal in this paper was to use the existing cache of list experiment and direct question prevalence estimates to characterize when such worries are warranted. 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 overreport turning out to vote, underreport participating in vote buying, underreport illegal behaviors like theft, and appear to report prejudice relatively accurately.

Our meta analysis also helps researchers to reason about the probable magnitude of sensitivity biases. We find that biases are typically smaller than five percentage points and are rarely as large as ten percentage points. While sensitivity bias is a real challenge for survey research, the extent of the problem is not as daunting as is sometimes presumed. It appears to be *normal* for respondents to accurately report their sensitive attitudes and behaviors.

From a research design perspective, sensitivity biases on the order of five to ten percentage points present scholars with a difficult choice. The high variance of list experiments means that, even if the assumptions hold, the substantive conclusions will be highly uncertain. In terms of root-mean-squared error, *unbiased* list experiments are dominated by *biased* direct questions at most sample sizes. In other words, unless sample sizes exceed 3,000 subjects, direct questions will tend to be more accurate than list experiments, even in the presence of non-negligible sensitivity biases.

These results should temper our expectations of sensitivity bias in many domains. In particular, we worry that scholars substitute their own norms of socially acceptable attitudes for the imagined norms of their subjects. For example, consider the relatively popular explanation for the 2016 US Presidential election polling miss that there were "Shy Trump Voters." Coppock (2017) reports the results of a large experiment that recovers no evidence of under-reporting support for Trump in any measured subgroup. That study leads us to suppose that the "Shy Trump" hypothesis caught because many in the media, polling organizations, and academia believed that Trump supports *should* be ashamed of themselves and supposing that they were leads to the expectation that direct reports were biased. In our view, expectations of sensitivity bias need to be rooted in a concrete theory of which reference group subjects are concerned with. In the candidate support case, one imagines that the relevant reference group is the polling organization. It appears that Trump supporters were not overly concerned with managing the impressions held by the presumably liberal polling organizations, perhaps because they were not ashamed of their support.

Survey subjects appear, on the whole, to cooperate with survey researchers. Conditional

on agreeing to participate in the survey, respondents seem to answer most direct questions with what they think is the truth. This paper aims to clarify under what conditions this baseline expectation is likely to be wrong. The clearest evidence of bias comes from domains like illegal activity or political support in violent settings, where the main concern of subjects is disclosure of dangerous information that would trigger truly negative consequences. In other domains, where the social reference group that we worry subjects are concerned with impressing is the survey team, we find less evidence of bias. This finding accords with recent experimental work showing limited evidence of experimenter demand effects (Mummolo and Peterson 2017; White et al. 2016).

Our study faces some important limitations. First and foremost, this is not a validation study. For most topics, we do not have access to the true prevalence rate. Indeed, this lack is what occasions the reliance of survey estimates of prevalence in the first place. Relatedly, this study relies on a comparison of two possibly flawed technologies. A statistically insignificant difference between the direct and list estimates of prevalence *might* indicate that the direct question does not exhibit sensitivity bias, but it might just as easily reflect similar biases in both methods. For example, if the list experiment does not protect responses from the reference group of the self, then no liars is violated and the list experiment is not unbiased for the true prevalence rate. The strong performance of list experiments on nonsensitive questions allays this concern to some degree, but not entirely.

Another limitation concerns the variability of the list experiment. As discussed in section 3.2, the power to detect moderate sensitivity bias, so our conclusion of limited bias in most direct measures may be an instance of "accepting the null of no bias" rather than properly failing to reject it. The more cautious and correct interpretation is that we can rule out biases greater than 10 or 15 percentage points in most cases. Biases on this order are of course very meaningful, but also difficult to detect with list experiments. Second, as noted, the statistical significance filter may bias our results. Moreover, only see list experiments that succeeded (i.e., that did not fail the design effects test) are likely to be available to us.

Despite the reasonable concern that the list experiment assumptions are unlikely to hold in at least some contexts, the technology appears to perform surprisingly well. In the 160 list experiments for which we have sufficient information, 156 pass the design effects test. In the 21 list experiments that are about putatively nonsensitive questions, 16 demonstrate no statistically significant difference between direct and list. The tool itself appears to generate approximately unbiased answers. At the same time, list experiments are known to be high variance; our analysis reveals that most list experiments that have been conducted thus

far have probably been too small for their research goals. Increasing sample sizes can of course be very expensive and researchers must weigh the data costs against the benefits of generating precise answers to their research questions, but the field should clearly move away from conducting underpowered list experiments. If the expected bias is on the order of 5 to 10 percentage points, direct questions (if ethically and operationally feasible) should be preferred if sample sizes are less than 3,000. Where possible, the design innovations described in section 3.5 should be incorporated.

Our summary result is that list experiments generally work as advertised, but they are not necessary in many settings. Researchers should ask themselves four questions when deciding whether to worry about the problem: is there a reference group respondents have in mind when answering, whom they believe can obtain their answers and for which respondents perceive a normative response to the question, and is there an expectation that the respondent (or others) will suffer costs if that normative response is not provided. In settings where list experiments are called for, researchers should combat the high variance of the list experiments the state-of-the-art designs and increase sample sizes far beyond those used in current practice.

	Number of Papers	Number of Studies
classroom	1	10
online	55	203
person	62	197
phone	19	31
self-report	5	13
	17	31
animal science	1	2
computer science	2	6
conservation biology	1	4
development	4	10
economics	1	1
environmental studies	4	13
management science	2	9
political science	101	276
psychology	8	50
public administration	1	12
public health	9	46
sociology	10	25
statistics	5	18
urban studies	1	3
veterinary medicine	1	1
	5	9
Obtained replication data set	43	107
Reconstructed dataset from tables in paper	32	63
Required information not available	30	91
Summary statistics reported in paper	58	224

Table 4: List Experiment Database Summary Statistics

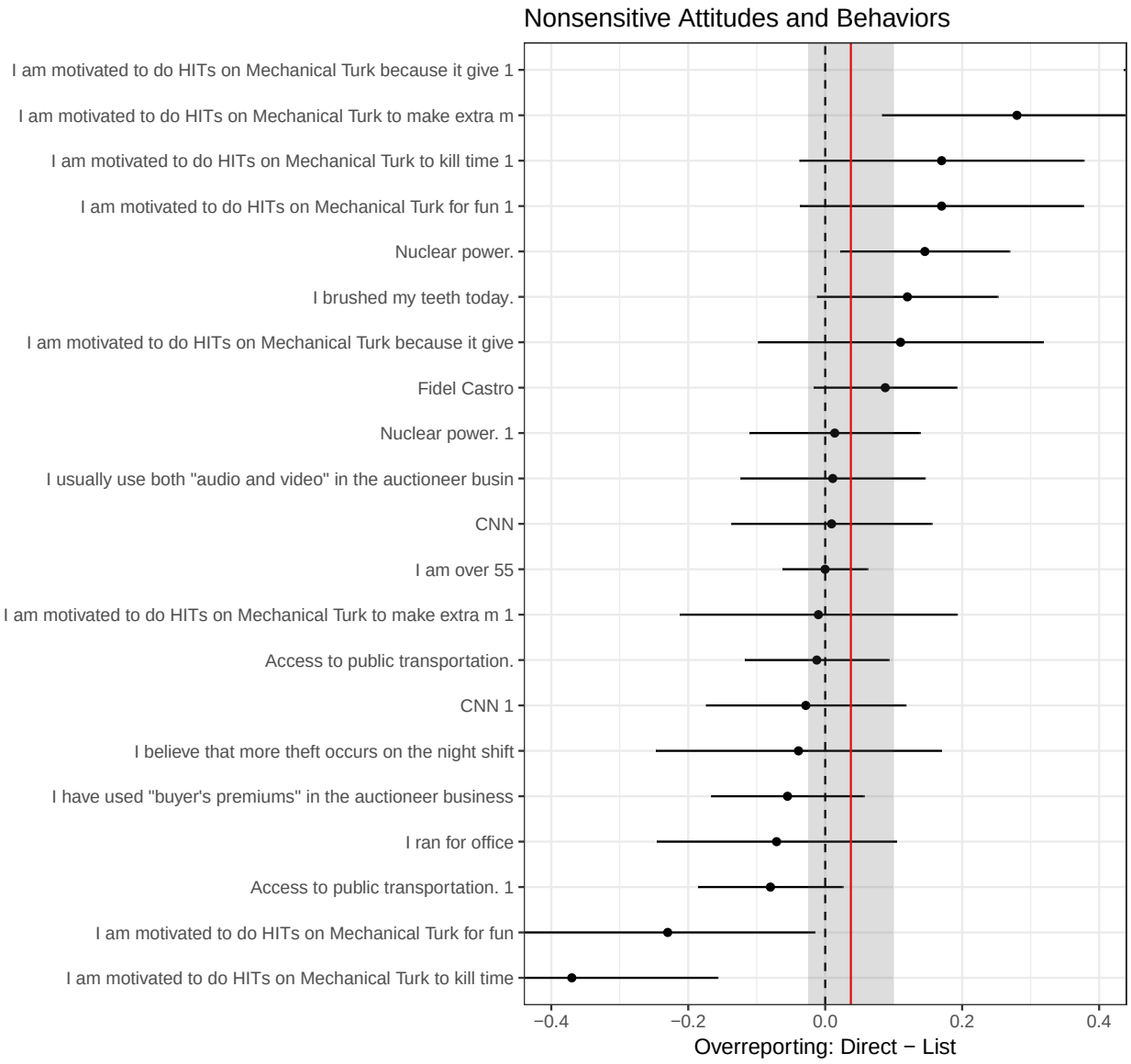

Figure 6: Estimates of Sensitivity Bias for Nonsensitive Questions

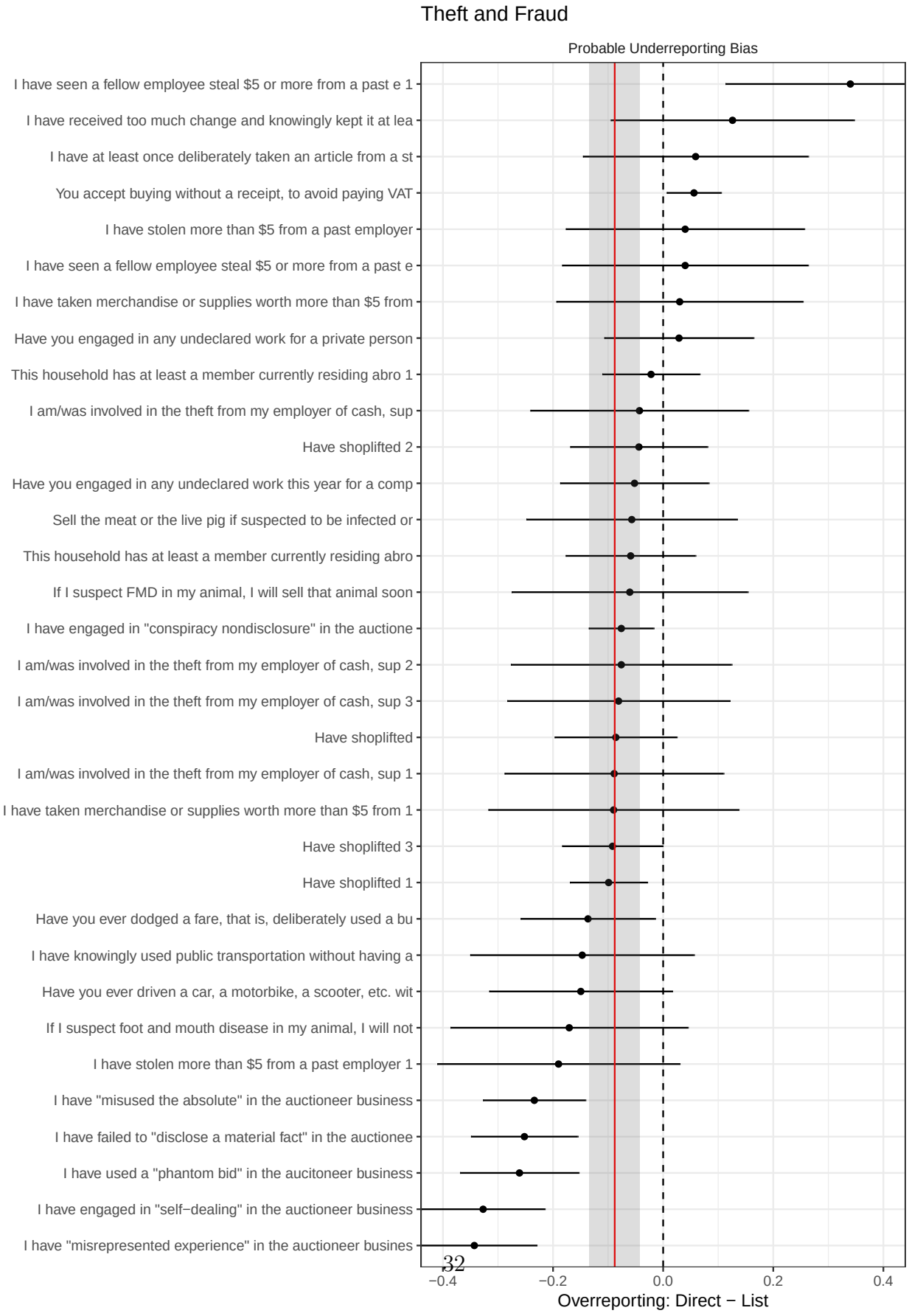

Figure 7: Estimates of Sensitivity Bias for Theft and Fraud

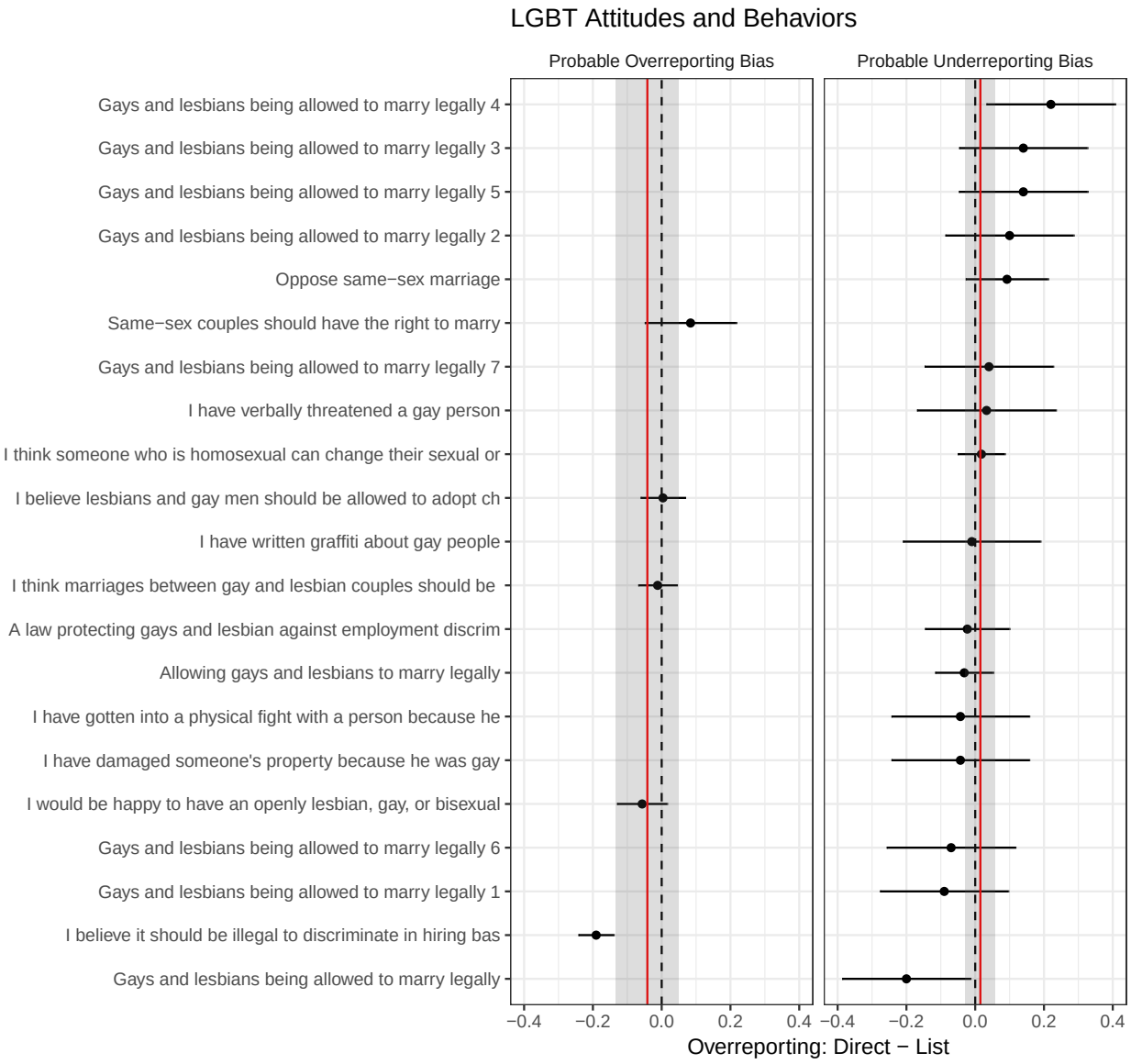

Figure 8: Estimates of Sensitivity Bias for LGBT attitudes

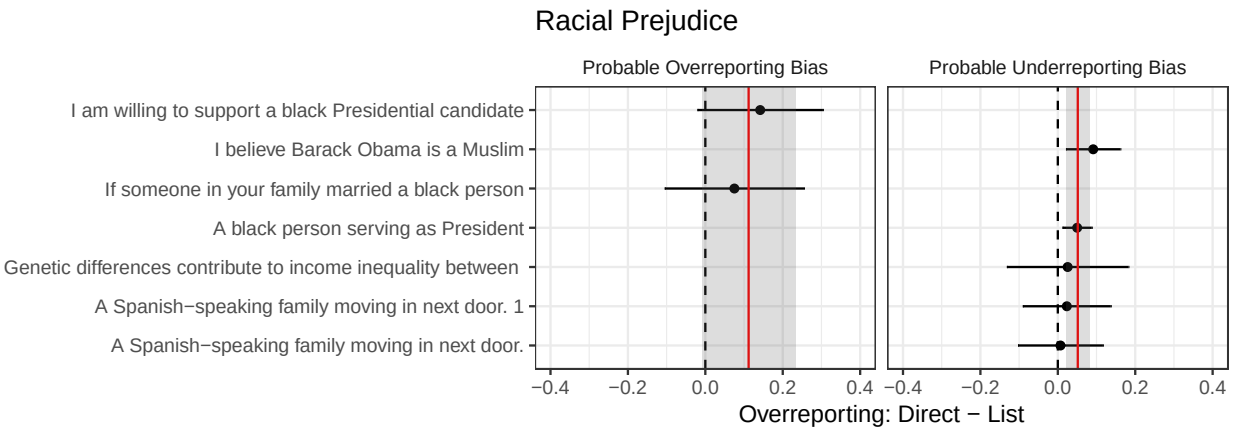

Figure 9: Estimates of Sensitivity Bias for Racial Prejudice

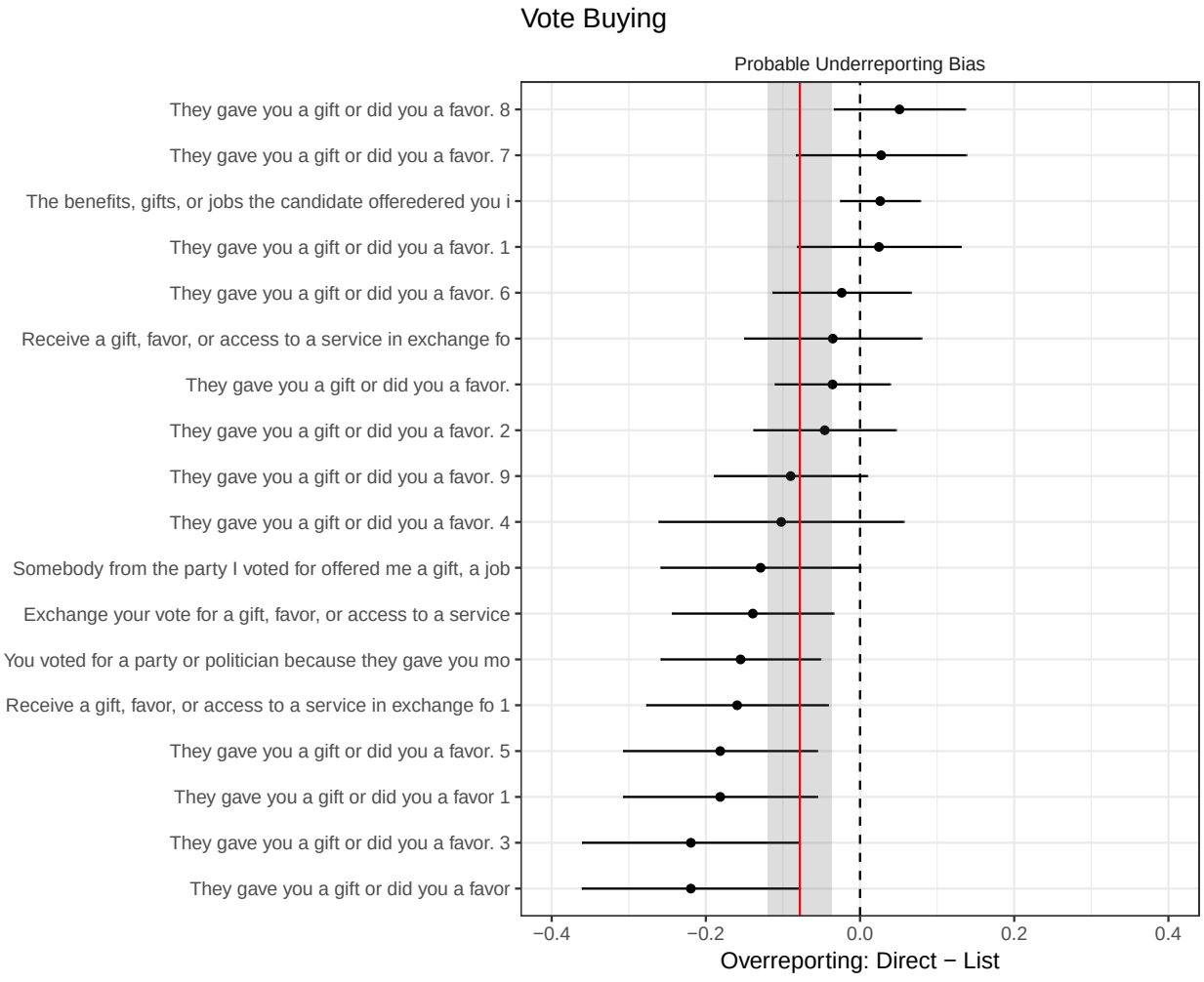

Figure 10: Estimates of Sensitivity Bias for Vote Buying

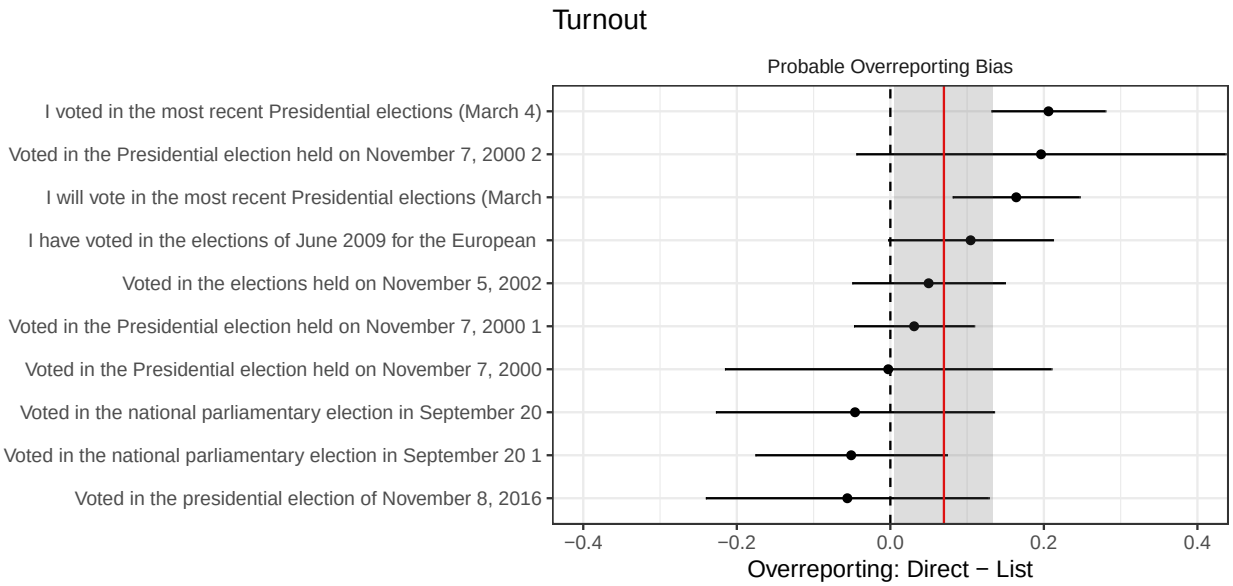

Figure 11: Estimates of Sensitivity Bias for Turnout

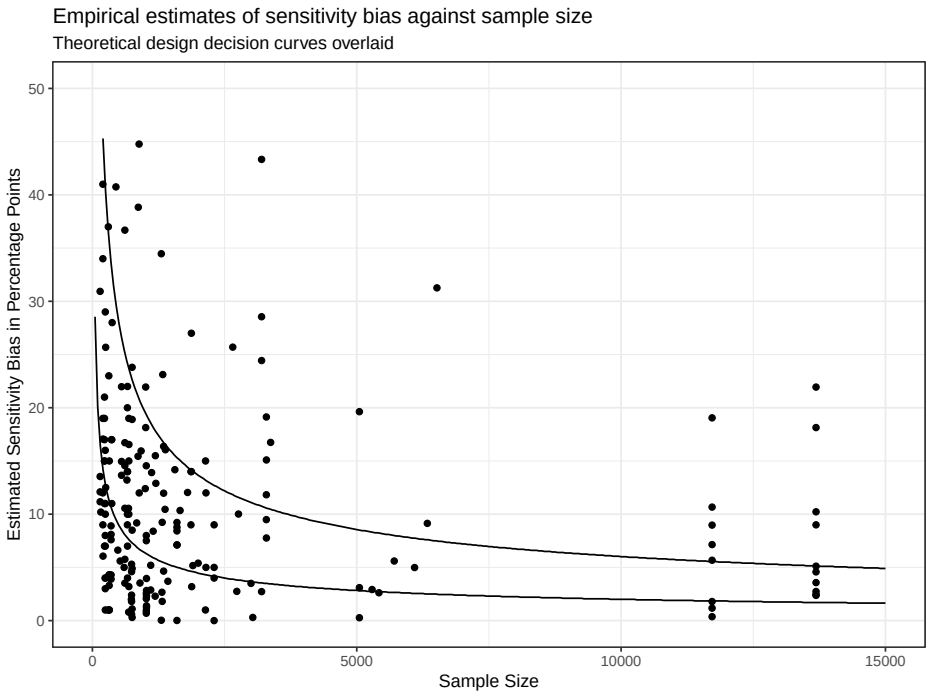

Figure 12: Empirical distribution of bias estimates by sample size

References

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

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.

Aquilino, William S., Debra L. Wright and Andrew J. Supple. 2000. "Response effects due to bystander presence in CASI and paper-and-pencil surveys of drug use and alcohol use." *Substance Use & Misuse* 35(6-8):845–867.

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.

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

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

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

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

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.

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.

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.

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

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

Deming, W. Edwards. 1944. "On errors in surveys." *American Sociological Review* 9(4):359–369.

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.

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.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Gerber, Alan and Neil Malhotra. 2008. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quarterly Journal of Political Science* 3(3):313–326.

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

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

Glennerster, Rachel and Kudzai Takavarasha. 2013. *Running randomized evaluations: A practical guide*. Princeton University 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.

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.

Hartmann, Petra. 1994. "Interviewing when the spouse is present." *International Journal of Public Opinion Research* 6(3):298–306.

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

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.

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

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

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

Krosnick, Jon A., Sowmya Narayan and Wendy R. Smith. 1996. "Satisficing in surveys: Initial evidence." *New Directions for Evaluation* (70):29–44.

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

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

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

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

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

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

Mummolo, Jonathan and Erik Peterson. 2017. "Demand Effects in Survey Experiments: An Empirical Assessment." Working paper, Department of Politics, Princeton University.

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.

Pollner, Melvin and Richard E. Adams. 1997. "The effect of spouse presence on appraisals of emotional support and household strain." *Public Opinion Quarterly* pp. 615–626.

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

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.

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

Tajfel, Henri and John C. Turner. 1979. "An integrative theory of intergroup conflict." *The social psychology of intergroup relations* 33(47):74.

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.

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

White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska and Connor Huff. 2016. "Investigator characteristics and respondent behavior in online surveys." *Journal of Experimental Political Science* 5(1):56–67.

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

Zipp, John F. and Joann Toth. 2002. "She said, he said, they said: The impact of spousal presence in survey research." *Public Opinion Quarterly* 66(2):177–208.

Appendix

A. Study results

polarity	category	est	se
Nonsensitive	Non-sensitive attitudes and behaviors	0.04	0.03
Probable Overreporting Bias	Other	0.06	0.02
Probable Overreporting Bias	Political Attitudes and Beliefs	0.13	0.03
Probable Overreporting Bias	Political Behavior	0.15	0.04
Probable Overreporting Bias	Prejudice	-0.04	0.05
Probable Overreporting Bias	Sexual Attitudes and Behaviors	0.08	0.04
Probable Overreporting Bias	Social Attitudes	0.19	0.03
Probable Underreporting Bias	Illegal Activity	-0.07	0.02
Probable Underreporting Bias	Other	-0.08	0.07
Probable Underreporting Bias	Personal Violence	-0.17	0.03
Probable Underreporting Bias	Political Attitudes and Beliefs	0.01	0.05
Probable Underreporting Bias	Political Behavior	-0.06	0.02
Probable Underreporting Bias	Prejudice	0.01	0.02
Probable Underreporting Bias	Sexual Attitudes and Behaviors	0.00	0.03
Probable Underreporting Bias	Socially Unacceptable Behavior	-0.00	0.03

Table 5: Meta-analytic summaries of overreporting, by polarity and category

B. Variance Calculations

We begin with the square of Eq. 3.4 in Gerber and Green (2012) to define the mean squared-error of the list experiment:

$$MSE_L = \frac{1}{N-1}\left\{\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m} + 2Cov(Y_i(0), Y_i(1))\right\}$$

Assume m = N/2

$$MSE_L = \frac{1}{N-1}\left\{Var(Y_i(0)) + Var(Y_i(1)) + 2Cov(Y_i(0), Y_i(1))\right\}$$

By no liars and no design effects, $Y_i(1) = Y_i(0) + p_i^*$

$$MSE_L = \frac{1}{N-1}\left\{Var(Y_i(0)) + Var(Y_i(0)+p_i^*) + 2Cov(Y_i(0), Y_i(0)+p_i^*)\right\}$$

Assume $Y_i(0) \perp p_i^*$ (and recalling that if $Y \perp Z$, then $cov(X, Y+Z) = cov(X,Y) + cov(X,Z)$)). Note that this assumption is *generous* to list experiments, because any positive correlation would increase list mse. Negative correlation in pos would violate design effects.

$$MSE_L = \frac{1}{N-1}\left\{Var(Y_i(0)) + Var(Y_i(0)) + Var(p_i^*) + 2 * Var(Y_i(0))\right\}$$

$$MSE_L = \frac{4Var(Y_i(0)) + Var(p_i^*)}{N-1}$$

Now, we define MSE_D. Recall that $p_i = 1$ if unit i answers yes when asked directly. Thus:

$$MSE_D = Var(p_i)/N + Bias^2$$

When is $MSE_L < MSD_D$?

$$\frac{4Var(Y_i(0))}{N-1} + \frac{Var(p_i^*)}{N-1} < \frac{Var(p_i)}{N} + Bias^2 \tag{1}$$

Define: $\pi*$ is true prevalence rate, p_i^* is true latent trait, p_i is observed direct q response. $W_i = p_i^* - p_i$. $Bias = E[W_i]$

$$\begin{aligned} Var(p_i) &= Var(p_i^* - W_i) \\ &= Var(p_i^*) + Var(W_i) + 2 * Cov(p_i^*, W_i) \end{aligned}$$

we need expression for $Cov(p_i^*, W_i)$

$$
\begin{aligned}
Cov(p_i^*, W_i) &= E[(p_i^* - \pi^*)(W_i - Bias)] \\
&= Bias * (1 - Bias - \pi^* + \pi^* * Bias) + \\
&\quad (\pi^* - Bias) * (\pi^* * Bias - Bias) + \\
&\quad (1 - \pi^*) * (\pi^* * Bias) \\
&= Bias - Bias * \pi^*
\end{aligned}
$$

plugging back in:

$$
\begin{aligned}
Var(p_i) &= Var(p_i^* - W_i) \\
&= Var(p_i^*) + Var(W_i) + 2 * Cov(p_i^*, W_i) \\
&= \pi^* * (1 - \pi^*) + Bias * (1 - Bias) + 2 * (Bias - Bias * \pi^*)
\end{aligned}
$$