Inter-American
Development Bank

Department of Research
and Chief Economist

TECHNICAL NOTE

No. IDB-TN-702

When to Protect?

Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior

Daniel W. Gingerich

November 2014

When to Protect?

Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior

Daniel W. Gingerich

Inter-American Development Bank

2014

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Gingerich, Daniel W., 1977-
When to protect? using the crosswise model to integrate protected and direct responses in surveys of sensitive behavior / Daniel W. Gingerich.
p. cm. — (IDB Technical Note ; 702)
Includes bibliographic references.
1. Human behavior—Statistical methods. 2. Quantitative research. I. Inter-American Development Bank. Department of Research and Chief Economist. II. Title. III. Series.
IDB-TN-702

http://www.iadb.org

Daniel W. Gingerich (dwg4c@virginia.edu)

Abstract

The applied social sciences have witnessed a growing use of sensitive survey techniques (SSTs) to study the relationship between facets of an individual's background and his propensity to engage in sensitive behavior. The rationale undergirding the use of these techniques is the assumption that the rate of misrepresentation and/or nonresponse under direct questioning among individuals bearing the sensitive trait would be so high as to make the use of direct questioning infeasible. But is this indeed the case? Presently available methodological tools do not provide an answer. The current paper presents a survey questioning strategy and corresponding statistical framework that simultaneously addresses the question of whether or not the use of a SST is required to study a given sensitive behavior, provides an estimate of the prevalence of the sensitive behavior in the population of interest, and, in its extended form, describes how individual characteristics relate to the likelihood of engaging in the behavior.

JEL Classification: C10, C51, C83
Keywords: Sensitive questions, Crosswise model, Randomized response, Expectation-maximization algorithm

1 Introduction

The challenge of characterizing the relationship between facets of an individual's background and his propensity to engage in sensitive forms of behavior is one that has long bedeviled social scientists. The need for empirical strategies to address this challenge has become especially acute in recent years, as an ever growing legion of researchers seeks to identify the fundamental predictors of unseemly but important phenomena such as corruption, vote buying, tax evasion, support for extremist movements, along with many other similarly sensitive objects of inquiry.

Recognizing both the potential of social surveys as well as the biases they invite when applied in standard form to sensitive issues, many scholars have begun to employ sensitive survey techniques (SSTs) in studies of sensitive behavior. Although there is variation in the format of such techniques, they all present the applied researcher with the same fundamental trade-off: greater protection of respondents, and, presumably, correspondingly lower bias due to legal and/or social desirability concerns against a loss of statistical efficiency due to the indirect manner in which the techniques query respondents about sensitive items. The rationale undergirding the use of these techniques in applied work is the assumption, often implicit, that the rate of misrepresentation and/or non-response under direct questioning among individuals bearing the sensitive trait would be so high as to make the bias-variance trade-off represented by SSTs well worth accepting in reasonably sized samples.

But is this indeed the case? Presently available methodological tools do not provide an answer. Typically, applied researchers employ SSTs to study a given sensitive topic because they have intuitions about the likely magnitude of evasive answer bias based on previous fieldwork with the target population, focus group sessions, or simple introspection. As reasonable and well informed as these intuitions may be, they are necessarily speculative and may be inaccurate in any particular setting.

The current paper presents a survey questioning strategy and corresponding statistical framework that simultaneously addresses the question of whether or not the use of a SST is required to study a given sensitive behavior, provides an estimate of the prevalence of the sensitive behavior in the population of interest, and, in its extended form, describes how individual characteristics relate to the likelihood of engaging in the behavior.

The questioning strategy developed in the paper is easy to describe. First, respondents are presented with a question about the sensitive behavior using a particular SST format. The paper considers the use of the so-called crosswise model, which provides anonymity via the commingling of responses about the sensitive behavior with responses about an innocuous question. (The crosswise model is mathematically identical to one version of the well-known randomized response technique, but it is administered in a different fashion; see below). Next, at a later stage of the survey respondents are queried directly about the same sensitive behavior, with the explicit

option of "choose not to respond directly" provided to them in case they deem a direct response to be uncomfortable or inappropriate. Observed responses about the sensitive behavior are thus a discrete combination of responses under the protection afforded by the SST and the absence of protection under direct questioning.

The statistical framework of the paper models the discrete response combinations using two key behavioral assumptions about the nature of misrepresentation under direct questioning along with a priori knowledge about the distribution of 'noise' intentionally introduced by the SST. The approach permits the applied researcher to estimate the probability of misrepresentation and non-response among bearers of the sensitive trait, thereby providing a principled basis for future consideration of the need to use SSTs to study the sensitive topic with similar populations.

Equally important, the approach harnesses the strengths of both survey formats in the sense that it incorporates all of the bias reducing advantages associated with the use of SSTs in high evasiveness settings while at the same time enhancing the precision of parameter estimates. Indeed, Monte Carlo simulations demonstrate that the performance of the paper's joint response model in terms of mean squared error is generally superior or equivalent to that of models based on direct or SST questioning only. Given that the addition of a direct question at the end of a survey instrument is virtually costless, this should be a compelling reason for many future studies employing SSTs to utilize a joint response approach as described here.

2 Related Methods

There are several existing approaches to studying the determinants of sensitive behaviors related to the one developed in this paper. One approach related to the current framework consists of a recently developed body of work on the use of item response theory (IRT) models with randomized response data (Böckenholt and van der Heijden, 2007; Böckenholt, Barlas, and van der Heijden, 2009; Fox, 2005; Fox and Wyrick, 2008; Fox, 2010 ch. 9). This work utilizes the randomized response technique to query respondents about attitudes or behaviors all thought to be reflective of a sensitive latent construct then employs an IRT measurement model to capture the relationship between the construct and the observed responses. In some instances, contributions to this literature have also specified structural models relating the individual characteristics of respondents to the sensitive latent construct. This approach has been used in applications that study the determinants of cheating by undergraduates (Fox and Meijer, 2008), consumer demand for pornography and prostitution (de Jong, Pieters, and Fox, 2010), and sexual attitudes (de Jong, Pieters, and Stremersch, 2012).

As in the current framework, efficiency advantages are obtained by harnessing responses to multiple questions about (related) sensitive phenomena. However, the conceptual goal of the current paper's framework is quite different than that of the aforementioned work. This paper

seeks not to combine survey responses to measure a continuous and inherently latent construct but rather to improve measurement of a binary outcome whose unobservability is assumed to stem only from the evasiveness of survey respondents under direct questioning.

A working paper more similar in spirit to the current study is Kraay and Murrell, 2013. This paper develops a framework for estimating the prevalence of sensitive behavior and candidness in surveys by utilizing direct questioning in conjunction with multiple randomized response questions. However, the paper assumes no difference in truthfulness across direct and SST questioning, and its core identifying assumption requires that the prevalence rate for all sensitive items be identical (irrespective of the content of those items). As such, the paper is necessarily silent about how question topic affects respondent evasiveness, an important concern for the applied researcher that is directly addressed by the framework developed here.

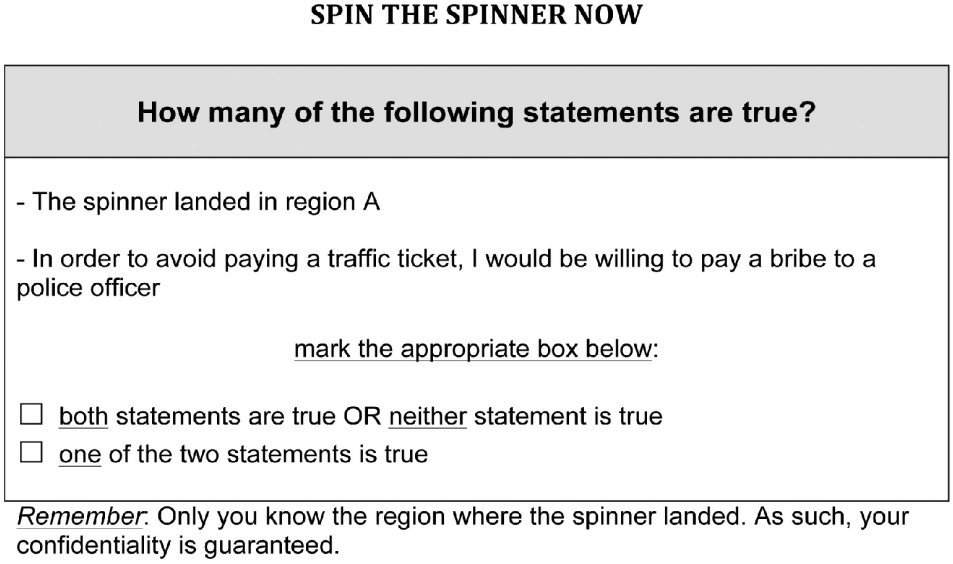
An alternative strategy for studying the determinants of sensitive attitudes related to the one developed here consists of the use of the item count technique (also referred to as the list experiment technique) (Miller, 1984). In recent years, a number of scholars have developed statistical methods based on item count data that are appropriate for studying the determinants of sensitive behavior (Corstange, 2009; Imai, 2011; Blair and Imai, 2012; Glynn, 2013). However, there appears to be only one published paper on the technique that develops a framework for combining item count responses with responses based on other question formats (Blair, Imai, and Lyall, forthcoming). This work combines item count responses with responses from endorsement experiments. Like the randomized response papers referenced above, the paper employs an underlying IRT measurement model to estimate a latent construct (e.g. support for insurgent movements) and to assess the role of explanatory variables in driving changes in value of the construct. Consequently, both the aim and the underlying statistical technology of that paper's framework are distinct from those outlined in the pages below.

3 Types of Sensitive Survey Techniques

There are two main sensitive survey techniques used in the social sciences: randomized response (RR) and the item count technique (ICT). Randomized response surveys query respondents about sensitive items by introducing a randomizing device, such as a spinner or a die, into the questioning process (Warner, 1965). More specifically, these surveys guarantee respondent confidentiality by requiring that a respondent's responses to a sensitive item be based not only upon the value of the sensitive attitude or behavior in question but also upon the realization of the randomizing device which she alone observes. Inference about the sensitive attitude or behavior proceeds by exploiting a priori information about the distribution of realizations generated by randomizing device.

Both meta analyses and validation studies have demonstrated the benefits associated with using randomized response instead direct questioning in studies of sensitive topics (Lensvelt-

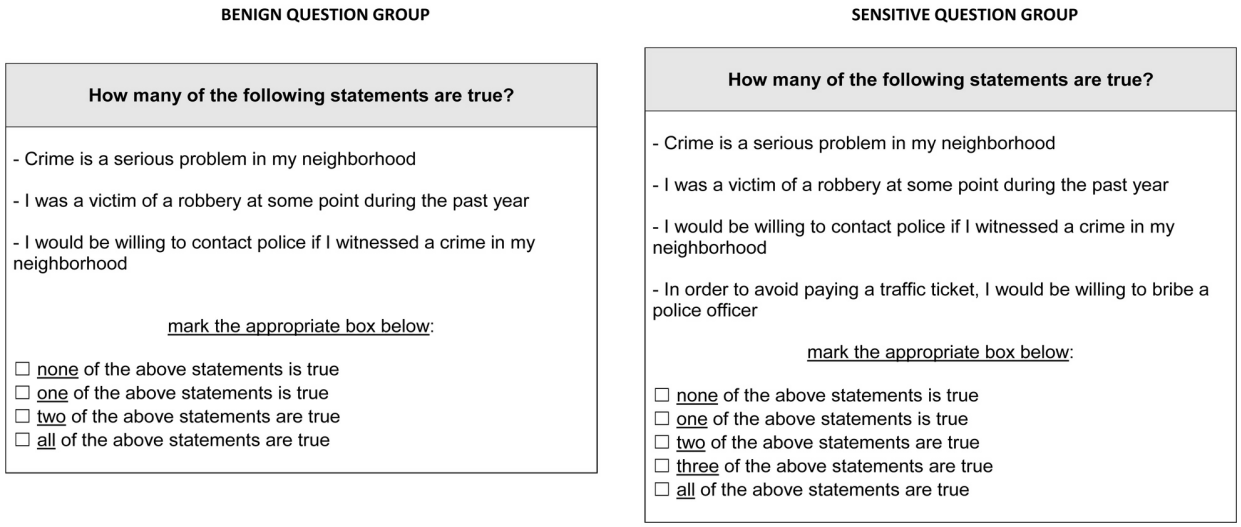
Figure 1. An example randomized response survey item



Mulders et al., 2005; Lamband and Stem, 1978; Tracy and Fox, 1981; van der Heijden et al., 2000; Fox, Avetisyan, and Palen, 2013). This track record and the convenient mathematical properties of the technique have helped spur a wave of applications in recent years, including investigations of the determinants of induced abortion in Mexico (Lara et al., 2006), social security fraud in the Netherlands (Lensvelt-Mulders et al., 2006), corruption within public bureaucracies in South America (Gingerich, 2010; Gingerich, 2013), the prevalence of xenophobia and anti-Semitism in Germany (Krumpal , 2012), as well as the role of anonymity on altruistic behavior in laboratory experiments (List et al., 2004; Moshagen, Hilbig, and Musch, 2011; Franzen and Pointner, 2012).

Figure 1 gives an example of a question asked using the Warner variant of the randomized response technique. The sensitive trait of interest is whether or not the respondent would be willing to pay a bribe to avoid a traffic ticket. The respondent is given a spinner divided into two regions A and B (with the probability that the spinner lands in region A not equal to 1/2). Prior to answering the question, she is directed to spin the spinner, observing the section where the spinner lands in private. Subsequently, the respondent is presented with two statements and asked how many are true. The first statement simply states that the spinner landed in region A. The second statement denotes a willingness to pay a bribe. The privacy of the respondent is protected by constraining the manner in which she is allowed to respond. In particular, there are only two potential responses: one response indicating that either both statements are true OR neither statement is true and another response indicating that only one of the two statements is true (but not specifying which is true). Since neither of the two responses necessarily indicates the possession or non-possession of the sensitive trait, the respondent's anonymity is guaranteed. For this reason, she may be liberated

Figure 2. An example item count technique survey item (benign and sensitive question groups)



from social desirability concerns that might otherwise prevent her from giving an honest answer about the trait.

Unlike randomized response (but similar to the crosswise model described below), the item count technique protects respondent confidentiality without using a randomizing device. Instead, respondent jeopardy is reduced by aggregating responses about the sensitive item with responses about a series of benign items. In the item count model, each individual in the sample is randomly assigned to one of two groups: a sensitive question group or a benign question group. In both groups, a given respondent is presented with a list of beliefs or activities and asked how many pertain to her. The two groups are presented with the same list of items save for one difference: the sensitive item is contained on the list for the sensitive group whereas it is omitted from the list for the benign question group. Inference proceeds through a comparison of the difference in average totals between the two groups. Due in part to its simplicity and ease of use, the item count technique has been widely applied in recent years to sensitive topics ranging from racial prejudice in the United States (Kuklinski et al., 1997; Gilens, Sniderman, and Kuklinski, 1998; Sniderman and Carmines, 1997), vote buying in Nicaragua (Gonzalez et al., 2012), attitudes towards immigration in the United States (Janus, 2010), corruption in foreign investment (Malesky, Gueorguiev, and Jensen, 2013), political patronage activities in Argentina (Oliveros, 2013), and interactions with drug trafficking organizations in Mexico (Magaloni et al., 2012).

Figure 2 provides an illustration of how the item count technique can be utilized to query respondents about willingness to bribe a police officer. Respondents in the benign question group are asked to denote how many of three statements about perceptions and experiences with crime are

true. Respondents in the sensitive question group are asked to denote how many of four statements are true, where the first three items are the same as in the benign group and the fourth is the sensitive statement about willingness to bribe a police officer. Since respondents simply divulge a sum total, the hope is that individuals in the sensitive group will feel comfortable in responding truthfully about their willingness to bribe.

Although RR and the ICT are useful workhorses for studying sensitive phenomena in surveys, they have a number of potential drawbacks. One potential drawback to randomized response may be its realm of applicability. In this regard, some scholars have suggested that the cognitive burden imposed by the use of a randomizing device can make RR difficult to use with populations that have very low levels of education (Böckenholt and van der Heijden, 2007). Indeed, a few studies have detected instances of non-compliance with the randomized response protocol, a problem which appears to be most common when the so-called forced response version of the randomized response technique is used (Edgell, Himmelfarb, and Duchan, 1982; Edgell, Duchan, and Himmelfarb, 1992; Azfar and Murrell, 2009).

There are also several drawbacks to the item count technique. One of these is the fact the technique requires investigators to collect two distinct samples, thereby increasing the logistical burden of the survey and reducing degrees of freedom for subsequent analysis. Another, arguably more important, drawback is the fact that the ICT provides incomplete protection to respondents. Individuals assigned to the sensitive question group who respond that all statements are true will directly reveal that they bear the sensitive trait. Moreover, respondents assigned to the sensitive group who wish to falsely signal that they do not bear the sensitive trait can easily and unambiguously do so by simply replying that none of the statements are true. Finally, it is somewhat challenging for investigators to control the level of respondent protection using the ICT, as this depends on quantities difficult to anticipate in advance of fielding a survey such as the frequencies and covariance of the responses to the benign statements.

In reaction to some of these concerns about RR and ICT, a recently developed body of work presents an alternative approach referred to as the crosswise model (Yu, Tian, and Tang, 2008; Tan, Tian, and Tang, 2009). The crosswise model is formally identical to the Warner variant of randomized response presented above, save for one difference: instead a employing a randomizing device to protect respondent confidentiality, the technique uses an indicator of membership in a non-sensitive group. The group indicator employed in the crosswise model is special in the sense that there are four conditions it must satisfy: 1) its value must be known to each respondent but unknown to survey administrators (and known by each respondent to be unknown to administrators); 2) it must be statistically independent of the sensitive trait of interest; 3) the proportion of individuals belonging to the group in the population of interest must be known in advance by investigators; and 4) the proportion of individuals belonging to the group must not be 1/2.

Figure 3. An example crosswise survey item

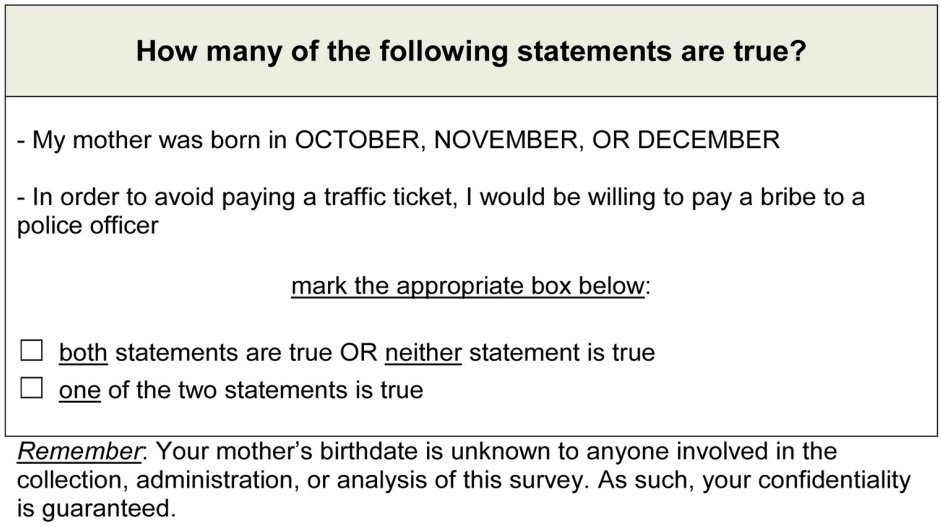


Figure 3 gives an example of a question asked using the crosswise model. As is apparent, the format of the question is nearly identical to that of the randomized response item. However, instead of having respondents condition their answers in part upon the privately observed realization of a randomizing device, the crosswise question prompts them to condition their answers upon the month of birth of their mother. In this case, nearly all respondents would know their own group assignment and they would also be aware that the survey enumerator did not know their group assignment (thereby ensuring privacy). Moreover, there is no realistic mechanism by the birth month of one's mother should be systematically tied to willingness to bribe, so the group indicator and the sensitive attitude would be independent of one another. Finally, the population proportion of membership in the two groups is verifiable based on census or actuarial records, and groups can easily be constructed such that the proportion of membership differs arbitrarily from $1/2$.

The virtues of the crosswise model include ease of implementation and a high-level of investigator control over the amount of protection afforded to respondents. Additionally, the technique requires no randomizing device nor splitting of the original sample, and it makes signalling behavior by respondents highly unlikely. Although the crosswise model is quite new, the empirical evidence that does exist on its effectiveness is favorable to its use (Jann, Jerke, and Krumpal 2012).

4 Integrating Protected and Direct Responses

Consider a setting in which each respondent i in a randomly selected sample of size n is first queried about her status on a sensitive trait $\theta \in \{0 \text{ ("absent")}, 1 \text{ ("present")}\}$ using the crosswise method (or equivalently, the Warner variant of the randomized response technique) then later asked if she would be willing to respond directly to a question about her status. If the respondent responds

affirmatively to the latter question, she is then prompted to directly divulge her true status on the sensitive trait. Interest resides in using the responses to these two questions to accomplish three goals: 1) calculate $\pi = \mathbb{E}[\theta_i] = \mathbb{P}(\theta_i = 1)$, the proportion of individuals who bear the trait of interest; 2) evaluate the returns to using the sensitive questioning technique to study the trait of interest in the target population; 3) estimate the influence of individual respondent characteristics on the incidence of the trait.

Notation. The combined response of respondent i to the two questions is denoted by the vector $Y_i = (y_i^D, y_i^A)$, where $y_i^D = \{0$ ("absent"), 1 ("present"), $\varnothing$ ("unwilling to respond directly")$\}$ is the observed response when i is queried about the sensitive trait directly and $y_i^A \in \{0, 1\}$ is the observed response when i is queried about the sensitive trait using the aforementioned sensitive question technique designed to guarantee anonymity. The observed response set is thus an array with six distinct elements, $\mathcal{Y} = \{(0,0), (0,1), (1,0), (1,1), (\varnothing, 0), (\varnothing, 1)\}$, with $k \in \mathcal{Y}$ representing an arbitrary element in this set. In the interest of notational compactness, the paper will henceforth use the simplification $Y_i \in \mathcal{Y} = \{1, 2, ...5, 6\}$, where each natural number $1, .., 6$ represents one of the six distinct response combinations. For the responses using the sensitive question technique, let $p \neq 1/2$ denote the probability that the first statement is true (e.g. the probability that the respondent's mother was born in the indicated interval of months for the crosswise model or that the spinner lands in region A for Warner RR). This quantity is known by the researcher prior to collecting the data.

Baseline Model

The paper's modeling strategy rests on two key assumptions. The first is called *honesty given protection*: given the protection afforded by the sensitive question technique, all respondents are assumed to respond as prompted by the technique (cf. Gingerich, 2010; Blair and Imai, 2012). Thus, if lying occurs in the survey responses, it is assumed to occur *only* when respondents are prompted to respond directly about the sensitive trait. This assumption is made by the vast majority of studies that employ sensitive question survey techniques. The second assumption is called *one-sided lying*: individuals who bear the sensitive trait may either lie about their status or refuse to respond when queried directly but those who do not bear the sensitive trait always either tell the truth or refuse to respond, they never falsely claim to bear the sensitive trait. Let λ_θ^T, λ_θ^L, and $1 - \lambda_\theta^T - \lambda_\theta^L$ denote the probability that, when queried directly, a respondent whose status is θ tells the truth about her status, lies about her status, or refuses to answer the question about her status, respectively. Formally, one-sided lying implies the parameter restriction $\lambda_0^L = 0$. The assumption follows naturally from the presumed direction of social desireability bias in sensitive surveys. If concerns about societal disapproval make it difficult for respondents bearing the sensitive trait to openly divulge their status, those same concerns should ensure that respondents not bearing the sensitive trait have no incentive to pass themselves off as bearers of the trait.

Table 1. Probability table for observed data under assumption of honesty given protection and one-sided lying

Y	Outcome	Probability	Frequency
1	$(y^D = 0, y^A = 0)$	$p\lambda_0^T(1-\pi) + (1-p)\lambda_1^L\pi$	n_1
2	$(y^D = 0, y^A = 1)$	$(1-p)\lambda_0^T(1-\pi) + p\lambda_1^L\pi$	n_2
3	$(y^D = 1, y^A = 0)$	$(1-p)\lambda_1^T\pi$	n_3
4	$(y^D = 1, y^A = 1)$	$p\lambda_1^T\pi$	n_4
5	$(y^D = \varnothing, y^A = 0)$	$p(1-\lambda_0^T)(1-\pi) + (1-p)(1-\lambda_1^T-\lambda_1^L)\pi$	n_5
6	$(y^D = \varnothing, y^A = 1)$	$(1-p)(1-\lambda_0^T)(1-\pi) + p(1-\lambda_1^T-\lambda_1^L)\pi$	n_6

Given these two assumptions, it is straightforward to characterize the probability of each combination of responses in the observed response set. Table 1 presents the relevant probability table. The formula presented in a given cell of the table expresses the probability of observing the particular response combination represented by that cell.

Let $I(.)$ be an indicator function equal to 1 if its argument is true, 0 otherwise, $\mathbb{P}_Y(k)$ be the probability of observing $Y_i = k$, and $\boldsymbol{\xi} = (\pi, \lambda_1^T, \lambda_1^L, \lambda_0^T)^\top$ be the full vector of parameters to be estimated. The likelihood function for the parameters given the observed responses is written:

$$L(\boldsymbol{\xi}|Y) = \prod_{i=1}^{n}\prod_{k=1}^{6}\mathbb{P}_Y(k)^{I(Y_i=k)}, \tag{1}$$

with the corresponding log-likelihood given by:

$$\ln L(\boldsymbol{\xi}|Y) = \sum_{k=1}^{6} n_k \ln \mathbb{P}_Y(k), \tag{2}$$

where $n_k = \sum_{i=1}^{n} I(Y_i = k)$ is the number of respondents exhibiting response category k.

Although applied researchers are typically focused on the estimation of π, the other parameters of the model are also of great substantive interest. These can be thought of as *diagnostic parameters*: they indicate the need (or lack thereof) to use a sensitive questioning technique to study the trait of interest in the target population. Particularly relevant is λ_1^T, the proportion of respondents bearing the sensitive trait who are willing to respond truthfully to a direct question about the trait. In a sense, the entire justification for utilizing a sensitive survey technique hinges on the value of this parameter. An estimated value of λ_1^T close to 1 indicates that the use of the sensitive questioning technique is unnecessary: researchers studying the same sensitive topic on the same population would have little to lose in bias and much to gain in statistical precision by using solely

direct questioning in future surveys. On the other hand, an estimated value of λ_1^T substantially below 1 indicates the importance of the respondent protection provided by the sensitive question technique. In particular, a value of λ_1^T well below 1 implies that substantial bias is incurred by querying respondents directly about the trait. Researchers studying the same sensitive topic on the same population would likely need to continue using the sensitive survey technique in the future.[1]

4.1 Modeling the Influence of Respondent Characteristics

More and more, social scientists' employ sensitive questioning techniques not simply to calculate prevalence estimates but rather to improve understanding of the factors that drive sensitive behaviors and attitudes. To this end, one can straightforwardly modify the model above in order to permit estimation of the influence of respondent characteristics on the sensitive outcome of interest.

To set up an explanatory model for the sensitive trait, one simply replaces the unconditional expectation parameter π with an appropriate parameterized conditional expectation function,

$$\pi_i = f(\mathbf{X}_i; \boldsymbol{\beta}) \tag{3}$$

where $\mathbf{X}_i$ is a vector of background characteristics and a constant, $\boldsymbol{\beta}$ is a parameter vector, and $f : \mathbb{R} \longmapsto [0, 1]$. A convenient choice for f is an inverse logit specification, $\pi_i = (1+\exp(-\mathbf{X}_i^\top \boldsymbol{\beta}))^{-1}$, although with continuous covariates and a sufficiently large sample alternative specifications of the linear predictor employing basis expansions and/or smoothing functions for $\mathbf{X}_i$ may also be an option.

Incorporating $\boldsymbol{\beta}$ into the full parameter vector $\boldsymbol{\xi}$ in place of π, the log-likelihood of the observed responses is now written:

$$\ln L(\boldsymbol{\xi}|Y, \mathbf{X}) = \textstyle\sum_{i=1}^{n} \sum_{k=1}^{6} I(Y_i = k) \ln \mathbb{P}_Y(k|\mathbf{X}_i), \tag{4}$$

where $\mathbb{P}_Y(k|\mathbf{X}_i)$ is the probability that respondent i's observed response is in category k given her background characteristics, the model for observed responses (e.g. the probabilities presented in Table 1), and the model for the conditional expectation of the sensitive trait.

4.2 Estimation and Inference

EM algorithm. The Expectation-Maximization (EM) algorithm provides a natural vehicle for attaining the maximum likelihood estimates (MLEs) of the parameters of interest. The algorithm is typically applied in settings that can be characterized as incomplete-data problems, where direct maximum likelihood estimation is made challenging by the absence of data in a standard format

[1] The parameter λ_0^T has a similar interpretation, although it seems unlikely that patterns of truthfullness (versus non-response) among those without the sensitive trait would vary as much as patterns of truthfulness among those with the trait.

(cf. Dempster, Laird, and Rubin, 1977). For crosswise or randomized response data, the incompleteness of the observed data structure stems from the fact that respondents' true values on the sensitive trait are not observed directly. Indeed, it is precisely the mixing of responses about the sensitive trait with responses about innocuous items (e.g. a relative's birthday, the realization of the spinner) that provides respondents with protection. For the responses generated by direct questioning, the incompleteness of the observed data structure stems from the fact that respondents may lie or not respond. Again, the challenge is that respondents' true values on the sensitive trait are not observed directly: some observed responses are truthful, others are misrepresentations, and others still are missing. Of course, estimates of the model parameters could be obtained fairly easily *if* one happened to be privy to the true value of the sensitive trait for each of the respondents. The paper's use of the E-M algorithm proceeds from this insight. In essence, the strategy is to recast the estimation problem from one in which all outcomes are known and fixed but for which the log-likelihood has a rather complicated form, to one in which only the probability of (at least some component of) the outcomes is known but for which the log-likehood is simpler to work with.

The E-M algorithm consists of several steps. The first is for the analyst to define an unobservable outcome Z, which, were it observable, would facilitate the estimation of MLEs. Once this has been accomplished, one characterizes the so-called complete-data log-likelihood function, $\ln L_c(\boldsymbol{\xi}|Y, Z)$, which is the log-likelihood that could be composed if both the actually observed and the unobservable data were observed. The subsequent step is to initialize the algorithm by choosing starting values for the parameters to be estimated, i.e. by setting $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)}$, where $\boldsymbol{\xi}^{(0)}$ are the starting values. Once starting values have be selected, one must complete the expectation step (E-step) of the algorithm, which requires the calculation of the quantity

$$Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{(0)}) = \mathbb{E}[\ln L_c(\boldsymbol{\xi}|Y, Z, \boldsymbol{\xi}^{(0)})]. \tag{5}$$

The $Q(.)$ function above is the expected value of the complete-data log-likelihood, evaluated at $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)}$ and taking as given the observed responses Y (and potentially the background characterstics, $\mathbf{X}$). After the current conditional value of the complete-data log-likelihood has been calculated, one proceeds to the maximization step (M-step) of the algorithm. This entails finding $\boldsymbol{\xi}^{(1)}$, which is the value of $\boldsymbol{\xi}$ that solves

$$\max_{\boldsymbol{\xi}} Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{(0)}). \tag{6}$$

Alternatively, if the above maximization problem is analytically intractible, one may choose $\boldsymbol{\xi}^{(1)}$ such that $Q(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(0)}) \geq Q(\boldsymbol{\xi}^{(0)}, \boldsymbol{\xi}^{(0)})$, in so doing defining a so-called generalized EM algorithm. In either case, once $\boldsymbol{\xi}^{(1)}$ has been obtained, the E-step and M-step are repeated with $\boldsymbol{\xi} = \boldsymbol{\xi}^{(1)}$. The algorithm then continues iterating through the E- and M-steps until convergence is achieved.

Table 2. Probability table for complete-data under assumption of honesty given protection and one-sided lying

Z	Y	Outcome	Probability	Expected Frequency
1	1	$(y^D = 0, y^A = 0, \theta = 0)$	$p\lambda_0^T(1-\pi)$	n_1'
2	1	$(y^D = 0, y^A = 0, \theta = 1)$	$(1-p)\lambda_1^L\pi$	n_1''
3	2	$(y^D = 0, y^A = 1, \theta = 0)$	$(1-p)\lambda_0^T(1-\pi)$	n_2'
4	2	$(y^D = 0, y^A = 1, \theta = 1)$	$p\lambda_1^L\pi$	n_2''
5	3	$(y^D = 1, y^A = 0, \theta = 1)$	$(1-p)\lambda_1^T\pi$	n_3
6	4	$(y^D = 1, y^A = 1, \theta = 1)$	$p\lambda_1^T\pi$	n_4
7	5	$(y^D = \varnothing, y^A = 0, \theta = 0)$	$p(1-\lambda_0^T)(1-\pi)$	n_5'
8	5	$(y^D = \varnothing, y^A = 0, \theta = 1)$	$(1-p)(1-\lambda_1^T-\lambda_1^L)\pi$	n_5''
9	6	$(y^D = \varnothing, y^A = 1, \theta = 0)$	$(1-p)(1-\lambda_0^T)(1-\pi)$	n_6'
10	6	$(y^D = \varnothing, y^A = 1, \theta = 1)$	$p(1-\lambda_1^T-\lambda_1^L)\pi$	n_6''

Baseline Model. Suppose, contrary to fact, that in addition to observing y_i^D and y_i^A for each respondent we could also observe θ_i, the sensitive trait of interest. Under this scenario, the outcome data for a given respondent would consist of 3×1 vector (y_i^D, y_i^A, θ_i). As a consequence of this expanded outcome space, the set of potential response combinations increases from six elements to ten. Let $Z_i \in \mathcal{Z} = \{1, 2, .., 10\}$ denote respondent i's unobservable outcome, where each natural number $1, .., 10$ represents one of the ten distinct response combinations. Table 2 presents the response combinations and probability table for this so-called complete data.

Let us begin by specifying the E-step of the E-M algorithm. To do so, one must first characterize the expected value of the log-likelihood of the complete data. Using Table 2, this quantity (ignoring an additive constant) can be written as:

$$\begin{aligned}\mathbb{E}[\ln L_c(\boldsymbol{\xi}|Y, Z)] = (n_A + n_D + n_E)\ln\pi + (n_B + n_C)\ln(1-\pi) + n_A \ln\lambda_1^T \\ + n_B \ln\lambda_0^T + n_C \ln(1-\lambda_0^T) + n_D \ln\lambda_1^L + n_E \ln(1-\lambda_1^T-\lambda_1^L),\end{aligned} \tag{7}$$

where $n_A = n_3 + n_4$, $n_B = n_1' + n_2'$, $n_C = n_5' + n_6'$, $n_D = n_1'' + n_2''$, and $n_E = n_5'' + n_6''$. Thus, the sufficient statistics for $\boldsymbol{\xi}$ are represented by the vector $S = (n_A, n_B, n_C, n_D, n_E)$, which, in turn, depends on the expected value of the unobserved frequencies.

For any model parameter $\xi \in \boldsymbol{\xi}$, let $\xi^{(j)}$ denote its value at the jth iteration of the E-M algorithm. The expected value of the unobserved frequencies can be expressed as:

$$
\begin{aligned}
n_1'^{(j)} &= n_1 \mathbb{P}(\theta_i = 0 | Y_i = 1)^{(j)} = n_1 \cdot \frac{p\lambda_0^{T(j)}(1-\pi^{(j)})}{p\lambda_0^{T(j)}(1-\pi^{(j)}) + (1-p)\lambda_1^{L(j)}\pi^{(j)}} \\
n_2'^{(j)} &= n_2 \mathbb{P}(\theta_i = 0 | Y_i = 2)^{(j)} = n_2 \cdot \frac{(1-p)\lambda_0^{T(j)}(1-\pi^{(j)})}{(1-p)\lambda_0^{T(j)}(1-\pi^{(j)}) + p\lambda_1^{L(j)}\pi^{(j)}} \\
n_5'^{(j)} &= n_5 \mathbb{P}(\theta_i = 0 | Y_i = 5)^{(j)} = n_5 \cdot \frac{p(1-\lambda_0^{T(j)})(1-\pi^{(j)})}{p(1-\lambda_0^{T(j)})(1-\pi^{(j)}) + (1-p)(1-\lambda_1^{T(j)}-\lambda_1^{L(j)})\pi^{(j)}} \\
n_6'^{(j)} &= n_6 \mathbb{P}(\theta_i = 0 | Y_i = 6)^{(j)} = n_6 \cdot \frac{(1-p)(1-\lambda_0^{T(j)}))(1-\pi^{(j)})}{(1-p)(1-\lambda_0^{T(j)})(1-\pi^{(j)}) + p(1-\lambda_1^{T(j)}-\lambda_1^{L(j)})\pi^{(j)}}
\end{aligned}
\tag{8}
$$

with $n_k''^{(j)} = n_k - n_k'^{(j)}$ for all $k \in \{1, 2, 5, 6\}$. Given the above, the current conditional sufficient statistics are calculated by plugging the expressions above into S, which produces the vector $S^{(j)} = (n_A, n_B^{(j)}, n_C^{(j)}, n_D^{(j)}, n_E^{(j)})$.

The M-step of the algorithm requires us to calculate the complete-data maximum-likelihood estimates of the model parameters at any given iteration of the algorithm. Maximization of the expected complete-data log-likelihood produces the following current conditional MLEs, $\widetilde{\boldsymbol{\xi}}^{(j+1)}$:

$$
\begin{aligned}
\widetilde{\pi}^{(j+1)} &= \frac{n_A + n_D^{(j)} + n_E^{(j)}}{n_A + n_D^{(j)} + n_E^{(j)} + n_B^{(j)} + n_C^{(j)}} \\
\widetilde{\lambda}_1^{T(j+1)} &= \frac{n_A}{n_A + n_D^{(j)} + n_E^{(j)}} \\
\widetilde{\lambda}_1^{L(j+1)} &= \frac{n_D^{(j)}}{n_A + n_D^{(j)} + n_E^{(j)}} \\
\widetilde{\lambda}_0^{T(j+1)} &= \frac{n_B^{(j)}}{n_B^{(j)} + n_C^{(j)}}.
\end{aligned}
\tag{9}
$$

Parameter estimates are obtained by cycling between $S^{(j)}$ and $\widetilde{\boldsymbol{\xi}}^{(j+1)}$ until convergence has been achieved.

Insight regarding the manner in which the statistical model utilizes information from the joint pattern of responses across question formats can be gleaned from an examination of the sufficient statistics contained in the expected complete-data log-likelihood. Consider first how the statistical model characterizes the unobserved frequency with which the sensitive trait is held in the sample. The first component of this frequency is n_A, the number of respondents who bear the sensitive trait and are willing to respond truthfully to a direct question about it. This quantity

is simply equal to the total number of respondents who respond in the affirmative to the direct question about the sensitive trait, i.e. $n_A = \sum_i^n y_i^D$. The equality is a direct consequence of the *one-sided lying* assumption. Since individuals who do not bear the sensitive trait never falsely claim that they do, any instance in which $y_i^D = 1$ is treated as an instance in which $\theta_i = 1$. In this way, the statistical model assumes that the number of respondents who bear the sensitive trait is never smaller than n_A. The second component of the frequency is n_D, the expected number of respondents who bear the sensitive trait but who lie in response to the direct question. It is equal to the sum of two products: the product of the number of respondents who respond $Y_i = (y_i^D = 0, y_i^A = 0)$ and the conditional probability of bearing the sensitive trait given this response pattern plus the product of the number of respondents who respond $Y_i = (y_i^D = 0, y_i^A = 1)$ and the conditional probability of bearing the sensitive trait given this latter response pattern. Defined analogously is the third component, n_E, the expected number of respondents who bear the sensitive trait but who choose not to respond when asked directly about the sensitive trait. The expected number of respondents who bear the sensitive trait is equal to the sum of the three aforementioned quantities, $n_A + n_D + n_E$.

Now consider the unobserved frequency with which the sensitive trait is not held. This quantity has two components. The first is n_B, the expected number of respondents who do not bear the sensitive trait and are willing to respond truthfully to a direct question about it. Since *one-sided lying* implies that these individuals never lie about their status, this quantity captures the expected number of respondents not bearing the sensitive trait who will not avoid answering the direct question. As above, this quantity is equal to the sum of two products: the product of the number of respondents who respond $Y_i = (y_i^D = 0, y_i^A = 0)$ and the conditional probability of not bearing the sensitive trait given this response pattern plus the product of the number of respondents who respond $Y_i = (y_i^D = 0, y_i^A = 1)$ and the conditional probability of not bearing the sensitive trait given this latter response pattern. The second component, defined analogously, is n_C, the expected number of respondents who do not bear the sensitive trait and who choose not to respond when asked directly about the sensitive trait. The expected number of respondents who do not bear the sensitive trait is equal to the sum, $n_B + n_C$.

The MLEs of the model parameters follow intuitively from these expected frequencies (see the system of equations in (9)). The estimated proportion of individuals bearing the sensitive trait, $\widetilde{\pi}$, is equal to the expected frequency of respondents bearing the trait divided by the total number of respondents. The estimated probability of a truthful response under direct questioning for individuals bearing the sensitive trait, $\widetilde{\lambda}_1^T$, is equal to the number of respondents who state that they have the trait under direct questioning divided by the expected frequency of respondents bearing the trait. The estimated probability of an untruthful response under direct questioning for those bearing the sensitive trait, $\widetilde{\lambda}_1^L$, is equal to the expected frequency of respondents who bear the sensitive trait but

lie about it under direct questioning divided by the expected frequency of respondents bearing the sensitive trait. Finally, the estimated probability of a truthful response under direct questioning for those not bearing the sensitive trait, $\widetilde{\lambda}_0^T$, is equal to the expected frequency of respondents who do not bear the sensitive trait and are willing to respond to direct questioning divided by the expected frequency of respondents not bearing the sensitive trait.

The conditional probabilities upon which the unobserved frequencies depend are a function of the unknown model parameters, the assumption of *honesty given protection*, and the known distribution of responses to the innocuous question utilized in the sensitive survey technique. The conditional probabilities assign a probability of bearing or not bearing the sensitive trait for each possible combination of observed response profiles. As can be seen by referencing the system of equations in (8) and the expressions in Table 2, they do this by employing Bayes' Rule. More specifically, the probability of the (unknown) trait status given an observed response profile is calculated as the ratio of the probability of the combination of the trait status and the observed response profile to the total probability of the observed response profile.

Modeling the Influence of Respondent Characteristics. To describe estimation for the model with respondent characteristics, it is necessary to introduce some additional notation. Let $\mathbb{P}_Z(z)_i = \mathbb{P}[Z_i = z|Y_i, \mathbf{X}_i, \boldsymbol{\xi}]$ be the conditional probability of an unobservable outcome $z \in \mathcal{Z}$ given the observed response Y_i, covariate vector $\mathbf{X}_i$, and the parameter vector $\boldsymbol{\xi}$. The expression $\mathbb{P}_Z(\{z_1, z_2\})_i = \mathbb{P}[Z_i = z_1 \cup Z_i = z_2|Y_i, \mathbf{X}_i, \boldsymbol{\xi}]$ will be similarly used to denote the conditional probability of the realization of either $Z_i = z_1$ or $Z_i = z_2$, where $z_1, z_2 \in \mathcal{Z}$.

Starting with the E-step, the expected value of the log-likelihood of the complete data is equal to

$$\begin{aligned}
&\mathbb{E}[\ln L_c(\boldsymbol{\xi}|Y, \mathbf{X}, Z)] \qquad (10)\\
&= \underbrace{\textstyle\sum_i^n \mathbb{P}_Z(\{2,4,5,6,8,10\})_i \ln f(\mathbf{X}_i;\boldsymbol{\beta}) + \sum_i^n \mathbb{P}_Z(\{1,3,7,9\})_i \ln(1 - f(\mathbf{X}_i;\boldsymbol{\beta}))}_{\text{binary regression component of log-likelihood}}\\
&+ \underbrace{\begin{aligned} &n_A \ln \lambda_1^T + \textstyle\sum_i^n \mathbb{P}_Z(\{1,3\})_i \ln \lambda_0^T + \sum_i^n \mathbb{P}_Z(\{7,9\})_i \ln(1 - \lambda_0^T)\\ &\qquad + \textstyle\sum_i^n \mathbb{P}_Z(\{2,4\})_i \ln \lambda_1^L + \sum_i^n \mathbb{P}_Z(\{8,10\})_i \ln(1 - \lambda_1^T - \lambda_1^L)\end{aligned}}_{\text{categorical component of log-likelihood}}
\end{aligned}$$

As can be seen above, the expected value of of the complete-data log-likelihood can be separated into two distinct components: a binary regression component that depends only on the parameter vector $\boldsymbol{\beta}$ and a categorical component that depends only on λ_1^T, λ_0^T, and λ_1^L.

The current conditional sufficient statistics for the full parameter vector $\boldsymbol{\xi}$ in this setting are equal to the expected values of the unobservable outcomes for each respondent in the sample given

her observed responses and background characteristics. On the jth iteration of the E-M algorithm, these are filled in as follows:

$$\mathbb{P}_Z(\{1,3\})_i^{(j)} = \begin{cases} 0 & \text{if} \quad Y_i \in \{3,4,5,6\} \\ \frac{p\lambda_0^{T(j)}(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))}{p\lambda_0^{T(j)}(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))+(1-p)\lambda_1^{L(j)}f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})} & \text{if} \quad Y_i = 1 \\ \frac{(1-p)\lambda_0^{T(j)}(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))}{(1-p)\lambda_0^{T(j)}(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))+p\lambda_1^{L(j)}f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})} & \text{if} \quad Y_i = 2 \end{cases}$$

$$\mathbb{P}_Z(\{2,4\})_i^{(j)} = \begin{cases} 0 & \text{if} \quad Y_i \in \{3,4,5,6\} \\ 1 - \mathbb{P}_Z(\{1,3\})_i & \text{if} \quad Y_i \in \{1,2\} \end{cases}$$

$$\mathbb{P}_Z(\{5,6\})_i^{(j)} = \begin{cases} 0 \quad \text{if} \quad Y_i \in \{1,2,5,6\} \\ 1 \quad \text{if} \quad Y_i \in \{3,4\} \end{cases}$$

$$\mathbb{P}_Z(\{7,9\})_i^{(j)} = \begin{cases} 0 & \text{if} \quad Y_i \in \{1,2,3,4\} \\ \frac{p(1-\lambda_0^{T(j)})(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))}{p(1-\lambda_0^{T(j)})(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))+(1-p)(1-\lambda_1^{T(j)}-\lambda_1^{L(j)})f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})} & \text{if} \quad Y_i = 5 \\ \frac{(1-p)(1-\lambda_0^{T(j)}))(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))}{(1-p)(1-\lambda_0^{T(j)})(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))+p(1-\lambda_1^{T(j)}-\lambda_1^{L(j)})f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})} & \text{if} \quad Y_i = 6 \end{cases}$$

$$\mathbb{P}_Z(\{8,10\})_i^{(j)} = \begin{cases} 0 & \text{if} \quad Y_i \in \{1,2,3,4\} \\ 1 - \mathbb{P}_Z(\{7,9\})_i & \text{if} \quad Y_i \in \{5,6\} \end{cases}$$

where, due to the disjointness of the unobservable outcomes, one can write:

$$\begin{aligned} \mathbb{P}_Z(\{2,4,5,6,8,10\})_i^{(j)} &= \mathbb{P}_Z(\{2,4\})_i^{(j)} + \mathbb{P}_Z(\{5,6\})_i^{(j)} + \mathbb{P}_Z(\{8,10\})_i^{(j)} \\ \mathbb{P}_Z(\{1,3,7,9\})_i^{(j)} &= 1 - \mathbb{P}_Z(\{2,4,5,6,8,10\})_i^{(j)}. \end{aligned}$$

The M-step of the algorithm is facilitated by the separability of the two components of the complete-data log-likelihood. Maximizing the expected complete-data log-likelihood with respect to the diagnostic parameters gives

$$\begin{aligned} \overset{\sim}{\lambda}_1^{T(j+1)} &= \frac{n_A}{n_A + \sum_i^n \mathbb{P}_Z(\{2,4\})_i^{(j)} + \sum_i^n \mathbb{P}_Z(\{8,10\})_i^{(j)}} \\ \overset{\sim}{\lambda}_1^{L(j+1)} &= \frac{\sum_i^n \mathbb{P}_Z(\{2,4\})^{(j)}}{n_A + \sum_i^n \mathbb{P}_Z(\{2,4\})_i^{(j)} + \sum_i^n \mathbb{P}_Z(\{8,10\})_i^{(j)}} \\ \overset{\sim}{\lambda}_0^{T(j+1)} &= \frac{\sum_i^n \mathbb{P}_Z(\{1,3\})_i^{(j)}}{\sum_i^n \mathbb{P}_Z(\{1,3\})_i^{(j)} + \sum_i^n \mathbb{P}_Z(\{7,9\})_i^{(j)}}. \end{aligned} \tag{11}$$

Since there is no closed-form solution for the parameters that maximize the log-likelihood of traditional binary regression models, one can calculate the current conditional parameter vector $\boldsymbol{\beta}^{(j+1)}$ by employing a Newton-Raphson step. This is justified by the global concavity of the log-likelihood when f consists of a logit or probit specification. More specifically, the global concavity of the log-likelihood in these circumstances ensures that by using a Newton-Raphson step in place of a global maximizer, one nonetheless has $Q(\boldsymbol{\xi}^{(j+1)}, \boldsymbol{\xi}^{(j)}) \geq Q(\boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}^{(j)})$, thereby guaranteeing convergence to the MLEs. For a logistic specification, for instance, one can write:

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbb{P}_Z(\{2,4,5,6,8,10\})_i^{(j)} - f(\mathbf{X}_i; \boldsymbol{\beta}^{(j)})) \tag{12}$$

where $f(\mathbf{X}_i; \boldsymbol{\beta}^{(j)}) = (1 + \exp(-\mathbf{X}_i^\top \boldsymbol{\beta}^{(j)}))^{-1}$ and $\mathbf{W}$ is an $n \times n$ diagonal matrix with ith element equal to $f(\mathbf{X}_i; \boldsymbol{\beta}^{(j)})(1 - f(\mathbf{X}_i; \boldsymbol{\beta}^{(j)}))$. As above, the algorithm proceeds by iterating through the sufficient statistics and parameter updates until convergence is achieved.

The explanatory model utilizes response data in a manner directly analogous to the baseline model, albeit with the distinction that the conditional probability of trait status given observed responses now varies across respondents as a function of individual characteristics. That is to say, respondents with identical observed response profiles are permitted to have different probabilities of bearing the sensitive trait based upon the underlying model of how respondent characteristics relate to trait status.

Quantities of Interest. Once the MLEs of the explanatory model have been estimated, there are a number of potentially useful quantities of interest that may be relevant to the applied researcher. Among these is the covariate-adjusted prevalence rate of the sensitive trait:

$$\widehat{\pi} = \frac{1}{n} \textstyle\sum_i^n f(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}). \tag{13}$$

Another quantity of interest may be the average predictive difference in the probability of bearing the sensitive trait associated with setting the value of target characteristic X_t equal to value v_1 instead of v_2:

$$\widehat{\Delta}_{v_1, v_2} = \frac{1}{n} \textstyle\sum_i^n f(X_{t,i} = v_1, \mathbf{X}_{l \neq t,i}; \widehat{\boldsymbol{\beta}}) - f(X_{t,i} = v_2, \mathbf{X}_{l \neq t,i}; \widehat{\boldsymbol{\beta}}). \tag{14}$$

Measures of Uncertainty. Measures of uncertainty such as confidence intervals and standard errors can be obtained via application of the parametric bootstap. Consider first confidence intervals. For the baseline model, these can be obtained by employing the following algorithm: 1) Draw a bootstrap sample b consisting of n iid draws of Y from the fitted multinomial density defined in Table 1, $Y_b \underset{\text{iid}}{\sim} \text{multinomial}(n; \widehat{\mathbb{P}}_Y(1), ..., \widehat{\mathbb{P}}_Y(6))$; 2) Apply the E-M algorithm to the bootstrap sample to calculate the bootstrap replicate of the parameter of interest; 3) Repeat this

process B times, where B is a large number (≥ 1000); 4) Calculate the lower (upper) bound of the $100(1-\alpha)$ percent confidence interval for the parameter as the $\alpha/2$ lower (upper) quantile of the bootstrap replicates. For the explanatory model, the first step of the algorithm is replaced by two additional steps: 1) Generate a bootstrap sample of the respondent characteristics, $\mathbf{X}_{ib}$, $i = 1, ..., n$, by drawing n observations with replacement from $\mathbf{X}$; 2) Generate a bootstrap sample of outcomes $Y_b = (Y_{1b}, ..., Y_{nb})^\top$ using the fitted multinomial density defined in Table 1 given the value of the bootstrap sample of characteristics, $Y_{ib} \sim$multinomial$(1; \widehat{\mathbb{P}}_Y(1|\mathbf{X}_{ib}), ..., \widehat{\mathbb{P}}_Y(6|\mathbf{X}_{ib}))$ for $i = 1, ..., n$. The remaining steps of the algorithm are then as specified above.

If instead of confidence intervals the full covariance matrix of $\boldsymbol{\xi}$ is desired, this quantity can be estimated by calculating the covariance matrix for the elements in the estimated parameter vector across the B bootstrap samples. Standard errors are equal to the square root of the diagonal entries of this covariance matrix.

5 Monte Carlo Analysis

In order to evaluate the performance of the paper's statistical model and estimation strategy, this section presents the results of a series of Monte Carlo simulations. The simulations were designed to gauge the performance of the explanatory model based on responses to both a sensitive question technique and direct questioning (the joint model) relative to a standard binary regression model that uses only responses generated by direct questioning (the DQ model) or a modified-binary regression model that uses only responses generated by a sensitive survey question technique (the SST model).

The simulations consider a setting in which there are two individual characteristics, X_{1i} and X_{2i}, responsible for variation in the sensitive behavior of interest. In the population, said characteristics are assumed to have a bivariate normal distribution with mean vector $(0.5, 1.5)$ and a covariance matrix with diagonal entries equal to $(0.1, 0.2)$ and off-diagonal entries equal to 0.1. The underlying relationship between the individual characteristics and the probability of having engaged in the sensitive behavior, π_i, is captured by an inverse logit function with linear predictor equal to $-1.0 + 2.0X_{1i} - 1.3X_{2i}$. Each Monte Carlo sample consists of a random draw from the population.

The simulations consider two distinct response scenarios for individuals who have and have not engaged in the sensitive behavior, a high evasiveness scenario and a low evasiveness scenario. In the high evasiveness scenario, the probability that an individual who has engaged in the sensitive behavior would respond truthfully to a direct question about the behavior is equal to 0.4, the probability that such an individual would lie under direct questioning is 0.4 (with corresponding non-response probability equal to 0.2), and the probability of non-response under direct questioning for an individual who had not engaged in the sensitive behavior is 0.2. In the low evasiveness

scenario, the probability that an individual who has engaged in the sensitive behavior would respond truthfully to a direct question about the behavior is equal to 0.7, the probability that such an individual would lie under direct questioning is 0.2 (with corresponding non-response probability equal to 0.1), and the probability of non-response under direct questioning for an individual who had not engaged in the sensitive behavior is 0.1.

In employing the simulations, the paper seeks to assess two things. Firstly, it seeks to verify the ability of the joint model elaborated in the text to recover the true values of all parameters. Secondly, it seeks to compare the performance of the joint model with those of the other two strategies mentioned above by examining differences in bias and mean squared error (MSE) for the parameters that all the models share, namely, those in the explanatory component, $\boldsymbol{\beta} = (-1.0, 2.0, -1.3)$. Performance is examined in this way for sample sizes of 2500 and 5000 respondents, respectively.

In evaluating the performance of the DQ model, coefficients were estimated based on complete case analysis, meaning that observations from individuals who provided no response were removed. Moreover, in order to ensure comparability of these results with those from the joint model, in the evaluation of the SST model the coefficients of interest were also estimated using an E-M algorithm. The details of this algorithm are provided in the appendix.

Table 3 presents the results of the Monte Carlo simulations. The first point to notice about the table is that, as anticipated, the parameter estimates produced by the joint model are centered on their true values, both for the diagnostic parameters as well as the parameters of the explanatory component of the model. A second crucial aspect of the table concerns the bias of the parameter estimates produced by the DQ model. These exhibited biases of various magnitudes, which, not surprisingly, reached fairly extreme levels under the high evasiveness scenario. Indeed, in the high evasiveness setting the estimator of the intercept in the DQ model was centered around a value nearly twice as small as the true value of the intercept.

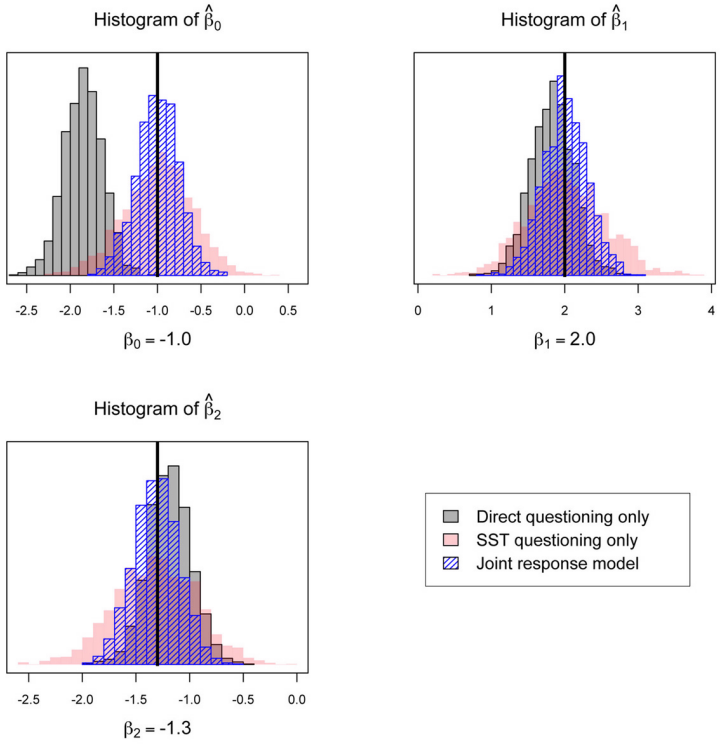
Like the parameter estimates from the joint model, those produced by the SST model were also centered around the true values of the parameters (a direct consequence of the assumption of honesty given protection). However, the variability of those estimates was *much* greater than that produced using the joint model, especially when the assumed level of evasiveness was low. For the coefficients on the individual characteristics, for instance, the MSE using the joint model was slightly more three times smaller than the MSE using the SST model under the high evasiveness scenario, and more than five times smaller under the low evasiveness scenario. Importantly, the attractive MSE performance of the joint model was not limited to a comparison with the SST model. Even in the low evasiveness scenario, where the advantages to protecting respondents through SST questioning are the weakest, the MSE performance of the joint model was either superior (for the intercept) or equivalent (for coefficients on characteristics) to the MSE performance of the direct model. In this sense, utilizing the joint model appears akin to a 'free lunch.' In certain scenarios,

Table 3. Comparisons of bias and mean squared error across alternative estimators (Monte Carlo simulations)

High Evasiveness Scenario												
	$\lambda_1^T = 0.400$		$\lambda_1^L = 0.400$		$\lambda_0^T = 0.800$		$\beta_0 = -1.000$		$\beta_1 = 2.000$		$\beta_2 = -1.300$	
n	2500	5000	2500	5000	2500	5000	2500	5000	2500	5000	2500	5000
	avg($\widehat{\lambda}_1^T$)		avg($\widehat{\lambda}_1^L$)		avg($\widehat{\lambda}_0^T$)		avg($\widehat{\beta}_0$)		avg($\widehat{\beta}_1$)		avg($\widehat{\beta}_2$)	
DQ	–	–	–	–	–	–	-1.852	-1.850	1.833	1.829	-1.193	-1.191
SQ	–	–	–	–	–	–	-1.017	-0.995	2.011	2.020	-1.307	-1.318
Joint	0.406	0.403	0.394	0.395	0.800	0.800	-1.003	-0.999	2.002	2.010	-1.308	-1.310
	MSE($\widehat{\lambda}_1^T$)		MSE($\widehat{\lambda}_1^L$)		MSE($\widehat{\lambda}_0^T$)		MSE($\widehat{\beta}_0$)		MSE($\widehat{\beta}_1$)		MSE($\widehat{\beta}_2$)	
DQ	–	–	–	–	–	–	0.827	0.771	0.189	0.116	0.094	0.053
SQ	–	–	–	–	–	–	0.292	0.155	0.611	0.286	0.289	0.141
Joint	0.003	0.002	0.005	0.003	<0.001	<0.001	0.117	0.059	0.181	0.091	0.089	0.047
Low Evasiveness Scenario												
	$\lambda_1^T = 0.700$		$\lambda_1^L = 0.200$		$\lambda_0^T = 0.900$		$\beta_0 = -1.000$		$\beta_1 = 2.000$		$\beta_2 = -1.300$	
n	2500	5000	2500	5000	2500	5000	2500	5000	2500	5000	2500	5000
	avg($\widehat{\lambda}_1^T$)		avg($\widehat{\lambda}_1^L$)		avg($\widehat{\lambda}_0^T$)		avg($\widehat{\beta}_0$)		avg($\widehat{\beta}_1$)		avg($\widehat{\beta}_2$)	
DQ	–	–	–	–	–	–	-1.323	-1.325	1.931	1.928	-1.256	-1.251
SQ	–	–	–	–	–	–	-1.017	-0.995	2.011	2.020	-1.307	-1.318
Joint	0.708	0.704	0.191	0.197	0.900	0.900	-0.998	-1.005	2.019	1.994	-1.315	-1.298
	MSE($\widehat{\lambda}_1^T$)		MSE($\widehat{\lambda}_1^L$)		MSE($\widehat{\lambda}_0^T$)		MSE($\widehat{\beta}_0$)		MSE($\widehat{\beta}_1$)		MSE($\widehat{\beta}_2$)	
DQ	–	–	–	–	–	–	0.167	0.137	0.110	0.055	0.053	0.027
SQ	–	–	–	–	–	–	0.292	0.155	0.611	0.286	0.289	0.141
Joint	0.007	0.004	0.008	0.004	<0.001	<0.001	0.079	0.041	0.117	0.055	0.054	0.027

Note: DQ denotes estimated logistic regression parameters based upon a complete case analysis of direct questioning responses only. SQ denotes logistic regression parameters based upon an analysis of sensitive question technique-based responses only (using an appropriately modified likelihood). Joint refers to logistic regression parameters based upon the joint response model developed in the paper. Two thousand random samples of size n were drawn for all Monte Carlo experiments. For the SQ and Joint models, we set $p = .25$.

Figure 4. Histograms of parameter estimates across alternative estimators (Monte Carlo Simulations)



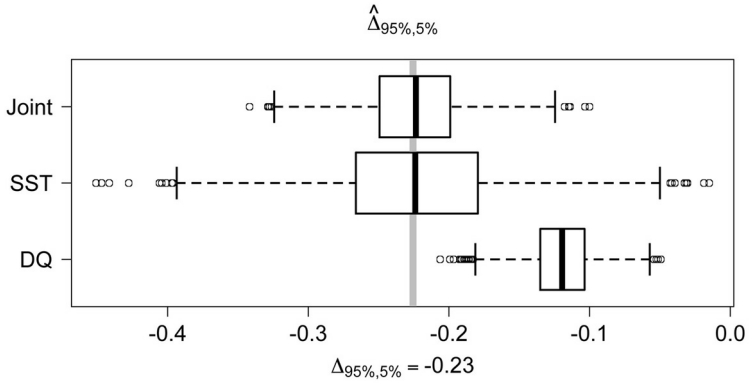
Note: The histograms shown above are for Monte Carlo simulations for the high evasiveness scenario with sample size n=5000. The thick vertical bars denote the true values of the parameters.

it is radically superior to either direct or SST questioning on their own, and under no scenario considered here does it fare worse than the aforementioned alternatives.

Figure 4 depicts the histograms of parameter estimates obtained for each parameter in the explanatory component of the models. As reported above, the histograms for parameters in the joint model and the SST model are both centered correctly, but the spread of the histograms for the joint model is significantly narrower throughout. On the other hand, the spread of the histograms of parameters in the joint model and DQ model are roughly equivalent, but the centering of the histograms is incorrect for all parameters in the DQ model, and radically so in the case of the intercept parameter.

Perhaps the best way to appreciate the performance of the joint model is to examine how it fares relative to the other approaches in estimating average predictive differences (APDs). Since coefficients in binary regression models typically have little substantive meaning on their own, APDs are often utilized by social scientists in binary regression settings as measures of the strength of the association between chosen explanatory variables and the outcomes of interest. Figure 5

Figure 5. Boxplots of average predictive differences across alternative estimators (Monte Carlo simulations)



Note: The boxplots shown above are for Monte Carlo simulations for the high evasiveness scenario with sample size $n = 5000$. The thick vertical bar denotes the true average predictive difference in the population.

presents boxplots of the estimated APDs associated with setting X_2 to the 95th percentile value in a given sample instead of the 5th percentile value. (Shown are the APDs for the high evasiveness scenario with a sample size of $n = 5000$.) In the simulation, the true population-level APD associated with such a shift is approximately 0.23. It is clearly seen that direct questioning produces highly biased estimates of the true APD: across the Monte Carlos, the average value of the estimated APD using the DQ model, -0.12, was nearly half as large as the true APD in absolute value. The APDs generated by the SST model and the Joint model were both unbiased. Nevertheless, the Joint model significantly outperformed the SST model in terms of precision, with a narrower interquartile range and a tighter distribution overall.

6 Conclusion

This paper has presented an intuitive joint response approach to modeling sensitive behavior. The approach utilizes responses about sensitive items both from indirect forms of questioning based upon sensitive survey techniques as well as upon responses based on direct survey questioning. In so doing, it allows applied researchers to perform three crucial tasks: 1) diagnose the need (or lack thereof) to use a SST to study the sensitive behavior of interest; 2) estimate the prevalence of the sensitive behavior; 3) estimate the relationship between the individual characteristics of respondents and the likelihood of engaging in the sensitive behavior.

One attractive feature of the approach is that it utilizes data from direct responses in a highly commonsensical way. In particular, the approach provides an estimate of the sensitive behavior of interest that is always greater than or equal to the proportion of respondents willing to admit under direct questioning that they have engaged in said behavior. This feature of the approach follows

directly from an assumption called one-side lying, which entails that individuals who have not engaged in the sensitive behavior never falsely claim that they do. While it may seem obvious that an estimation strategy designed to calculate the prevalence of sensitive behaviors should bound its estimates in this way, extant approaches based solely upon responses generated by sensitive survey techniques do *not* do so. This is an important point, since in practice it is certainly possible for prevalence estimates based solely on SSTs to be below those obtained via direct questioning.

While hopefully expanding the toolkit of social scientists interested in sensitive forms of behavior, it is important to recognize that this paper nevertheless leaves much to be done. An implicit assumption of the paper's framework is that the evasiveness of respondents under direct questioning is unaffected by whether or not they are previously asked about the sensitive item in SST format. In future work, the reasonableness of this assumption could be assessed by randomly assigning a small subset of survey respondents to receive the sensitive item in DQ format only, thereby permitting a comparison of responses to direct questioning when that format is the only one employed to responses to direct questioning when a joint response approach is utilized. Such a comparison would be valuable in assessing the diagnostic utility of the study's approach.

References

Azfar, O. and P. Murrell. 2009. "Identifying Reticent Respondents: Assessing the Quality of Survey Data on Corruption and Values." *Economic Development and Cultural Change* 57(2): 387-411.

Blair, G. and K. Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1): 47-77.

Blair, G., K. Imai, and J. Lyall. Forthcoming. "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan." *American Journal of Political Science*.

Böckenholt, U. and P. G. M. van der Heijden. 2007. "Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses." *Psychometrika* 72(2): 245-262.

Böckenholt, U., S. Barlas, and P. G. M. van der Heijden. 2009. "Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior." *Journal of Applied Econometrics, Special Issue: New Econometric Models in Marketing* 377–392.

Corstange, D. 2009. "Sensitive questions, truthful answers? Modeling the list experiment with LISTIT." *Political Analysis* 17: 54–63.

de Jong, M. G., R. Pieters, and J.P. Fox. 2010. "Reducing social desirability bias through item randomized response: An application to measure underreported desires." *Journal of Marketing Research* 47 (1): 14-27.

de Jong, M. G., R. Pieters, and S. Stremersch. 2012. "Analysis of sensitive questions across cultures: An application of multigroup item randomized response theory to sexual attitudes and behavior." *Journal of Personality and Social Psychology* 103(3): 543-564.

Dempster, A.P., N.M. Laird, and D. B Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1-38.

Edgell, S. E., K. L. Duchan, and S. Himmelfarb. 1992. "An empirical test of the unrelated question randomized response technique." *Bulletin of the Psychonomic Society* 30(2): 153-156.

Edgell, S. E., S. Himmelfarb, and K.L. Duchan. 1982. "Validity of forced responses in a randomized response model." *Sociological Methods & Research* 11(1): 89-100.

Fox, J. P. 2010. "Randomized Item Response Models (Chapter 9)" In *Bayesian item response modeling: Theory and applications*. Springer.

—. 2005. "Randomized item response theory models." *Journal of Educational and Behavioral statistics* 30(2): 189-212.

— and C. Wyrick. 2008. "A mixed effects randomized item response model." *Journal of Educational and Behavioral Statistics* 33(4): 389-415.

—— and R. R. Meijer. 2008. "Using item response theory to obtain individual information from randomized response data: An application using cheating data." *Applied Psychological Measurement* 32(8): 595-610.

——, M. Avetisyan, and J. Palen. 2013. Mixture randomized item-response modeling: a smoking behavior validation study. *Statistics in Medicine* 32(27): 4821-4837.

Franzen, A., and S. Pointner. 2012. "Anonymity in the dictator game revisited." *Journal of Economic Behavior & Organization* 81(1): 74-81.

Gilens, M., Sniderman, P. M., and Kuklinski, J. H. 1998. "Affirmative action and the politics of realignment." *British Journal of Political Science* 28 (1), 159-183.

Gingerich, D. W. 2013. *Political Institutions and Party-Directed Corruption in South America: Stealing for the Team*. Cambridge: Cambridge University Press.

——. 2010. "Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys." *Political Analysis* 18: 349-380.

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

Gonzalez-Ocantos, E., De Jonge, C. K., Meléndez, C., Osorio, J., and Nickerson, D. W. 2012. "Vote buying and social desirability bias: Experimental evidence from Nicaragua." *American Journal of Political Science* 56(1): 202-217.

Imai, K. 2011. "Multivariate regression analysis for the item count technique." *Journal of the American Statistical Association* 106: 407–16.

Jann, B., J. Jerke, and I Krumpal. 2012. "Asking sensitive questions using the crosswise model an experimental survey measuring plagiarism." *Public Opinion Quarterly* 76(1): 32-49.

Janus, A. L. 2010. "The influence of social desirability pressures on expressed immigration attitudes." *Social Science Quarterly* 91(4): 928-946.

Kraay, A., and P. Murrell. 2013. "Misunderestimating corruption." Policy Research Working Paper 6488, World Bank, Washington, DC

Krumpal, I. 2012. "Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning." *Social science research* 41(6): 1387-1403.

Kuklinski, J. H., P. M. Sniderman, K. Knight, T. Piazza, P.E. Tetlock, G.R. Lawrence, and B. Mellers. 1997. "Racial prejudice and attitudes toward affirmative action." *American Journal of Political Science* 41 (2): 402-419.

Lamb Jr, C. W., and D.E. Stem Jr. 1978. "An empirical validation of the randomized response technique." *Journal of Marketing Research* 15(4): 616-621.

Lara, D., S. G. García, C. Ellertson, C. Camlin, and J. Suarez. 2006. "The measure of induced abortion levels in Mexico using randomized response technique." *Sociological Methods*

and Research 35: 279–30.

Lensvelt-Mulders, G. J. L., J. J. Hox, P. G. M. van der Heijden, and C. J. M. Maas. 2005. "Meta-analysis of randomized response research: Thirty years of validation." *Sociological Methods and Research* 33: 319–48.

Lensvelt-Mulders, G. J. L., P. G. M. van der Heijden, O. Laudy, and G. van Gils. 2006. "A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security." *Journal of the Royal Statistical Society Series A* 169: part 2: 305–18.

List, J. A., R. P. Berrens, A. K. Bohara, and J. Kerkvliet. 2004. "Examining the role of social isolation on stated preferences." *American Economic Review* 94: 741-752.

Magaloni, B., A. Díaz-Cayeros, V. Romero, and A. Matanock. 2012. "The enemy at home: exploring the social roots of criminal organizations in Mexico." Unpublished paper.

Malesky, E. J., D. Gueorguiev, and N. M. Jensen. 2013. "Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment." Available at SSRN 1967670.

Miller, J. D. 1984. "A new survey technique for studying deviant behavior," PhD thesis, George Washington University, Department of Sociology.

Moshagen, M., B. E. Hilbig, and J. Musch. 2011. "Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games." *European Journal of Social Psychology* 41(5): 638-644.

Oliveros, V. 2013. "A Working Machine: Patronage Jobs and Political Services in Argentina." PhD thesis, Columbia University.

Sniderman, P. M., and E. G. Carmines. 1997. *Reaching Beyond Race*. Cambridge: Harvard University Press.

Tan, M. T., G. L. Tian, and M. L. Tang. 2009. "Sample surveys with sensitive questions: a nonrandomized response approach." *The American Statistician* 63(1).

Tracy, P. E., and J. A. Fox. 1981. "The validity of randomized response for sensitive measurements." *American Sociological Review* 46:187–200.

van der Heijden, P. G. M., G. van Gils, J. Bouts, and J. J. Hox. 2000. "A comparison of randomized response, computer assisted self interview and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit fraud." *Sociological Methods & Research* 28: 505–37.

Warner, S. L. 1965. "Randomized response: A survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60: 63–9.

Yu, J. W., G. L. Tian, and M. L. Tang. 2008. "Two new models for survey sampling with sensitive characteristic: design and analysis." *Metrika* 67(3): 251-263.

Appendix – EM Algorithm for SST Questioning Only

Presented here is the EM algorithm employed in the Monte Carlo analyses to estimate the parameters of the explanatory model when only SST questioning is utilized. Appendix Table 4 presents the relevant probability table for the complete data under the assumption of honesty given protection.

Appendix Table 4. Probability table for complete data for SST questioning only

Z	Y	Outcome	Probability
1	0	$(y^A = 0, \theta = 0)$	$p(1-\pi)$
2	0	$(y^A = 0, \theta = 1)$	$(1-p)\pi$
3	1	$(y^A = 1, \theta = 0)$	$(1-p)(1-\pi)$
4	1	$(y^A = 1, \theta = 1)$	$p\pi$

Utilizing the notation employed previously in the text, the expected value of the log-likelihood of the complete data (up to a constant) is equal to

$$\mathbb{E}[\ln L_c(\boldsymbol{\xi}|Y, \mathbf{X}, Z)] = \sum_i^n \mathbb{P}_Z(\{2,4\})_i \ln f(\mathbf{X}_i;\boldsymbol{\beta}) + \sum_i^n (1 - \mathbb{P}_Z(\{2,4\})_i) \ln(1 - f(\mathbf{X}_i;\boldsymbol{\beta})), \tag{A1}$$

where $\mathbb{P}_Z(\{2,4\})_i$ is the conditional probability of bearing the sensitive trait given a respondent's observed response and background characteristics.

On the jth iteration of the E-M algorithm, this quantity is equal to

$$\mathbb{P}_Z(\{2,4\})_i^{(j)} = \begin{cases} \frac{(1-p)f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})}{p(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))+(1-p)f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})} & \text{if} \quad Y_i = 0 \\ \frac{pf(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})}{pf(\mathbf{X}_i;\boldsymbol{\beta}^{(j)})+(1-p)(1-f(\mathbf{X}_i;\boldsymbol{\beta}^{(j)}))} & \text{if} \quad Y_i = 1 \end{cases} \tag{A2}$$

Based on the above, the current conditional parameter vector $\boldsymbol{\beta}^{(j+1)}$ is identical to that shown in (13), save for the fact that $\mathbb{P}_Z(\{2,4\})_i^{(j)}$ is inserted into the equation as the relevant response variable. The EM algorithm proceeds by iterating through $\boldsymbol{\beta}^{(j)}$ and $\mathbb{P}_Z(\{2,4\})_i^{(j+1)}$ until convergence is achieved.