Advance Access publication December 30, 2015

*Political Analysis* (2016) 24:132–156
doi:10.1093/pan/mpv034

# When to Protect? Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior

**Daniel W. Gingerich**
*Department of Politics, University of Virginia, Charlottesville, VA 22903*
*e-mail: dwg4c@virginia.edu (corresponding author)*

**Virginia Oliveros**
*Department of Political Science, Tulane University, New Orleans, LA 70118*

**Ana Corbacho and Mauricio Ruiz-Vega**
*Western Hemisphere Department, International Monetary Fund, Washington, DC 20431*

Edited by R. Michael Alvarez

Sensitive survey techniques (SSTs) are frequently used to study sensitive behaviors. However, existing strategies for employing SSTs lead to highly variable prevalence estimates and do not permit analysts to address the question of whether the use of an SST is actually necessary. The current article presents a survey questioning strategy and corresponding statistical framework that fills this gap. By jointly analyzing survey responses generated by an SST (the crosswise model) along with direct responses about the sensitive behavior, the article's framework addresses the question of whether the use of an SST is required to study a given sensitive behavior, provides an efficient estimate of the prevalence of the sensitive behavior, and, in its extended form, efficiently estimates how individual characteristics relate to the likelihood of engaging in the behavior. The utility of the approach is demonstrated through an examination of gender differences in proclivities towards corruption in Costa Rica.

## 1 Introduction

The challenge of characterizing the relationship between facets of an individual's background and her 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, and 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

*Authors' note*: This article was prepared when Ana Corbacho was Sector Economic Advisor and Daniel Gingerich and Virginia Oliveros were visiting scholars at the Inter-American Development Bank. The data, code, and any additional materials required to replicate all analyses in this article are available on the Political Analysis Dataverse of Harvard University's Dataverse Network at: http://dx.doi.org/10.7910/DVN/2AIHQF. Supplementary materials for this article are available on the *Political Analysis* Web site.

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

The questioning strategy developed in the article is easy to describe. First, respondents are presented with a question about the sensitive behavior using a particular SST format. In the article, we consider 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 [RR] 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 article 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 (MSE) is generally superior or equivalent to that of models based on direct or SST questioning only. This is likely to be of considerable practical importance in applications, as it is well known that SSTs can produce prevalence estimates that are highly variable. Given that the addition of a direct question at the end of a survey instrument is virtually costless, we believe that the statistical advantages of our framework offer 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 we develop in this article. One approach related to our framework consists of a recently developed body of work on the use of item response theory (IRT) models with RR data (Fox 2005; Böckenholt and van der Heijden 2007; Fox and Wyrick 2008; Böckenholt, Barlas, and van der Heijden 2009). This work utilizes the RR technique to query respondents about attitudes or behaviors all thought to be reflective of a sensitive latent construct and 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 our framework, efficiency advantages are obtained by harnessing responses to multiple questions about (related) sensitive phenomena. However, the conceptual goal of our framework is quite different from that of the aforementioned work. We seek 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 on RR that is similar in spirit to ours 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 RR 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.

Our strategy toward parameter estimation is to employ the Expectation–Maximization (EM) algorithm. In this respect, our article is similar to Bourke and Moran (1988), which describes estimation using RR data without covariates, and Blair, Imai, and Zhou (2015), which describes estimation using RR data with covariates. The latter paper provides a rich and broad-reaching presentation of statistical tools using RR data, covering topics such as power analysis, the use of RR as a predictor variable, and non-compliance with RR protocols. However, the article does not touch on the central issue of concern here, which is the joint modeling of direct and protected responses in studies of sensitive topics.

An alternative to RR for studying the determinants of sensitive attitudes consists of the use of the item count technique (ICT) (also referred to as the list experiment technique) (Miller 1984). 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). Moreover, in recent years, methods based on combining responses from the ICT with responses from other modalities of questioning have also begun to emerge.

For instance, Blair, Imai, and Lyall (2014) combine item count responses with responses from endorsement experiments. Like several of the RR papers referenced above, the article 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.

Aronow et al. (2015) develop a statistical framework for combining responses from the ICT with direct questioning. This article makes a similar assumption to ours about patterns of lying under direct questioning, and it also emphasizes the efficiency advantages of incorporating direct questioning. However, Aronow et al. (2015) focus exclusively on the task of estimating the prevalence of a sensitive trait. In addition to prevalence estimation, the framework of our article accommodates a multivariate explanatory model of the factors that drive trait status and it allows the user to explicitly estimate the rate of lying under direct questioning by respondents who hold the sensitive trait. Additionally, our framework is unique in that it adjusts for refusals and non-response under direct questioning, a commonly encountered pattern in surveys on sensitive topics.

## 3 Types of SSTs

There are two main SSTs used in the social sciences: RR and the ICT. RR 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 on the value of the sensitive attitude or behavior in question but also on 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 RR instead of direct questioning in studies of sensitive topics (Lamb and Stem 1978; Tracy and Fox 1981; van der Heijden et al. 2000; Lensvelt-Mulders et al. 2005; Fox, Avetisyan, and Palen 2013; Rosenfeld, Imai, and Shapiro 2014). This track record and the convenient mathematical properties of the technique have helped spur numerous 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, 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; Franzen and Pointner 2012).

Unlike RR (but similar to the crosswise model we describe below), the ICT 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 ICT 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), vote buying in Nicaragua (Gonzalez-Ocantos et al. 2012), and corruption in foreign investment (Malesky, Gueorguiev, and Jensen 2015), to interactions with drug trafficking organizations in Mexico (Magaloni et al. 2012).

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 RR 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 RR protocol, a problem which appears to be most common when the so-called forced response version of the RR technique is used (Edgell, Himmelfarb, and Duchan 1982; Azfar and Murrell 2009).

There are also several drawbacks to the ICT. One of these is the fact that 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 the 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 RR, save for one difference: instead of 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) its proportion in the population of interest must

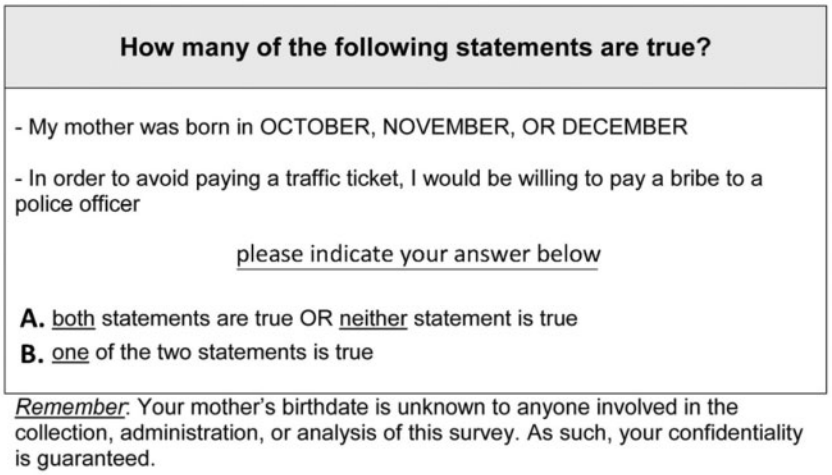


**Fig. 1** An example crosswise survey item.

be known in advance by investigators; and (4) the proportion of individuals belonging to the group must not be 1/2.

Figure 1 gives an example of a question asked using the crosswise model. The sensitive trait of interest is whether or not the respondent would be willing to pay a bribe to a police officer in order to avoid a traffic ticket. However, rather than asking respondents directly about willingness to bribe, the survey format presents respondents with two statements and asks how many are true. The first statement deals with membership in a non-sensitive group. It reads "My mother was born in OCTOBER, NOVEMBER, OR DECEMBER." The second statement denotes a willingness to pay a bribe in order to avoid a traffic ticket. 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 from social desirability concerns that might otherwise prevent her from giving an honest answer about the trait.

In this example, it should be clear that mother's birth month satisfies the conditions for a non-sensitive group indicator outlined above. Nearly all respondents would know their mother's birth month, and they would also be aware that the survey enumerator did not know this information (thereby ensuring privacy). Moreover, there is no realistic mechanism by which 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 frequency of births by month is typically verifiable based on census or actuarial records, and groups can easily be constructed such that the probability 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 signaling 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 $\theta \in \{0$ ("absent"), 1 ("present")$\}$ using the crosswise method

(or equivalently, the Warner variant of the RR 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. Our 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; and (3) estimate the influence of individual respondent characteristics on the incidence of the trait.

### 4.1 *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"), $\emptyset$ ("unwilling to respond directly")$\}$ is the observed response when $i$ is queried about the sensitive trait directly and $y_i^A \in \{0$ ("B"), 1 ("A")$\}$ 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, $\mathcal{Y} = \{(0, 0), (0, 1), (1, 0), (1, 1), (\emptyset, 0), (\emptyset, 1)\}$, with $k \in \mathcal{Y}$ representing an arbitrary element in this set. In the interest of notational compactness, we will henceforth use the simplification $Y_i \in \mathcal{Y} = \{1, 2, ...5, 6\}$, where each natural number $1, \ldots, 6$ 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). This quantity is known by the researcher prior to collecting the data.

### 4.2 *Baseline Model*

Our 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 majority of studies that employ sensitive question survey techniques. Our 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 $\lambda_\theta^T$, $\lambda_\theta^L$, and $1 - \lambda_\theta^T - \lambda_\theta^L$ denote the probability that, when queried directly, a respondent whose status is $\theta$ 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 desirability 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.

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

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

**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 | $(y^D = \emptyset, y^A = 0)$ | $p(1-\lambda_0^T)(1-\pi) + (1-p)(1-\lambda_1^T-\lambda_1^L)\pi$ | $n_5$ |
| 6 | $(y^D = \emptyset, y^A = 1)$ | $(1-p)(1-\lambda_0^T)(1-\pi) + p(1-\lambda_1^T-\lambda_1^L)\pi$ | $n_6$ |

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 $\pi$, 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 $\lambda_1^T$, 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 an SST hinges on the value of this parameter. An estimated value of $\lambda_1^T$ 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 $\lambda_1^T$ substantially below 1 indicates the importance of the respondent protection provided by the sensitive question technique. In particular, a value of $\lambda_1^T$ 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 SST in the future.[1]

### 4.3 *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 $\pi$ 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 $f : \mathbb{R} \mapsto [0, 1]$. 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.

[1]The parameter $\lambda_0^T$ has a similar interpretation, although it seems unlikely that patterns of truthfulness (versus non-response) among those without the sensitive trait would vary as much as patterns of truthfulness among those with the trait.

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

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

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.

### 4.4 *Estimation and Inference*

#### 4.4.1 EM algorithm

The EM algorithm provides a natural vehicle for attaining the maximum-likelihood estimates (MLEs) of the parameters of interest in our model. 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 (cf. Dempster, Laird, and Rubin 1977). For crosswise or RR 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 a 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* we happened to be privy to the true value of the sensitive trait for each of our respondents. Our use of the EM algorithm proceeds from this insight. In essence, our 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-likelihood is simpler to work with.

The EM 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, that is, by setting $\xi = \xi^{(0)}$, where $\xi^{(0)}$ are the starting values. Once starting values have been selected, one must complete the expectation step (E-step) of the algorithm, which requires the calculation of the quantity

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

The $Q(.)$ function above is the expected value of the complete-data log-likelihood, evaluated at $\xi = \xi^{(0)}$ and taking as given the observed responses $Y$ (and potentially the background characteristics, $\mathbf{X}$). 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 $\xi^{(1)}$, which is the value of $\xi$ that solves

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

Alternatively, if the above maximization problem is analytically intractable, one may choose $\xi^{(1)}$ such that $Q(\xi^{(1)}, \xi^{(0)}) \geq Q(\xi^{(0)}, \xi^{(0)})$, in so doing defining a so-called generalized EM algorithm. In either case, once $\xi^{(1)}$ has been obtained, the E-step and M-step are repeated with $\xi = \xi^{(1)}$. 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 | $(y^D = \emptyset, y^A = 0, \theta = 0)$ | $p(1-\lambda_0^T)(1-\pi)$ | $n_5'$ |
| 8 | 5 | $(y^D = \emptyset, y^A = 0, \theta = 1)$ | $(1-p)(1-\lambda_1^T-\lambda_1^L)\pi$ | $n_5''$ |
| 9 | 6 | $(y^D = \emptyset, y^A = 1, \theta = 0)$ | $(1-p)(1-\lambda_0^T)(1-\pi)$ | $n_6'$ |
| 10 | 6 | $(y^D = \emptyset, y^A = 1, \theta = 1)$ | $p(1-\lambda_1^T-\lambda_1^L)\pi$ | $n_6''$ |

### 4.4.2 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 $\theta_i$, the sensitive trait of interest. Under this scenario, the outcome data for a given respondent would consist of a $3 \times 1$ vector $(y_i^D, y_i^A, \theta_i)$. 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 $1, \ldots, 10$ represents one of the ten distinct response combinations. Table 2 presents the response combinations and probability table for this so-called complete data.

We begin by specifying the E-step of the EM algorithm. To do so, we 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 $j$th iteration of the EM 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 MLEs 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, $\tilde{\xi}^{(j+1)}$:

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

Parameter estimates are obtained by cycling between $S^{(j)}$ and $\tilde{\xi}^{(j+1)}$ 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, that is, $n_A = \sum_i^n y_i^D$. 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 equation (9)). The estimated proportion of individuals bearing the sensitive trait, $\tilde{\pi}$, 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, $\tilde{\lambda}_1^T$, 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, $\tilde{\lambda}_1^L$, 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, $\tilde{\lambda}_0^T$, 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 SST. 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 equation (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.

### 4.4.3 Modeling the influence of respondent characteristics

To describe estimation for the model with respondent characteristics, we must 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 $z_1, z_2 \in \mathcal{Z}$.

We begin with the E-step. The expected value of the log-likelihood of the complete data is equal to

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

As can be seen above, the expected value 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 $\lambda_1^T$, $\lambda_0^T$, and $\lambda_1^L$.

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 $j$th iteration of the EM algorithm, these are filled in as follows:

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

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

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

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

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

where due to the disjointness of the unobservable outcomes, we can write

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

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

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

Since there is no closed-form solution for the parameters that maximize the log-likelihood of traditional binary regression models, we 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, we nonetheless have $Q(\boldsymbol{\xi}^{(j+1)}, \boldsymbol{\xi}^{(j)}) \geq Q(\boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}^{(j)})$, thereby guaranteeing convergence to the MLEs. For a logistic specification, for instance, we have

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

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 $i$th 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 on the underlying model of how respondent characteristics relate to trait status.

### 4.4.4 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:

$$\hat{\pi} = \frac{1}{n}\sum_{i}^{n} f(\mathbf{X}_i; \hat{\boldsymbol{\beta}}). \tag{13}$$

Another quantity of interest may be the average predictive difference (APD) 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 $v_2$:

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

### 4.4.5 Measures of uncertainty

Measures of uncertainty such as confidence intervals and standard errors can be obtained via application of the parametric bootstrap. 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 $n$ i.i.d. draws of $Y$ from the fitted multinomial density defined in Table 1, $Y_b \underset{\text{iid}}{\sim} \text{multinomial}(n; \hat{\mathbb{P}}_Y(1), ..., \hat{\mathbb{P}}_Y(6))$; (2) apply the EM 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 ($\geq 1000$); and (4) calculate the lower (upper) bound of the $100(1 - \alpha)\%$ 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 $\mathbf{X}$; and (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, $Y_{ib} \sim \text{multinomial}(1; \hat{\mathbb{P}}_Y(1|\mathbf{X}_{ib}), ..., \hat{\mathbb{P}}_Y(6|\mathbf{X}_{ib}))$ 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 $\xi$ 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 our 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, $\pi_i$, 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.

Our 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, we are interested in assessing two things. First, we would like to verify the ability of the joint model elaborated in the text to recover the true values of all parameters. Second, we seek to compare the performance of the joint model with those of the other two strategies mentioned above by examining differences in bias and MSE for the parameters that all the models share, namely, those in the explanatory component, $\boldsymbol{\beta} = (-1.0, 2.0, -1.3)$. We examine performance in this way for sample sizes of 2500 and 5000 respondents, respectively.[2]

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.

[2]In evaluating the performance of the DQ model, we estimated coefficients based on complete case analysis, meaning that observations from individuals who provided no response were removed. Moreover, in order to ensure comparability of our results with the joint model, in our evaluation of the SST model we also estimated the coefficients of interest using an EM algorithm. The details of this algorithm are provided in the Online Appendix.

**Table 3** Comparisons of bias and MSE across alternative estimators (Monte Carlo simulations)

| | $\lambda_1^T = 0.400$ | | $\lambda_1^L = 0.400$ | | $\lambda_0^T = 0.800$ | | $\beta_0 = -1.000$ | | $\beta_1 = 2.000$ | | $\beta_2 = -1.300$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *High evasiveness scenario* | | | | | | | | | | | | |
| $n$ | 2500 | 5000 | 2500 | 5000 | 2500 | 5000 | 2500 | 5000 | 2500 | 5000 | 2500 | 5000 |
| | avg($\hat{\lambda}_1^T$) | | avg($\hat{\lambda}_1^L$) | | avg($\hat{\lambda}_0^T$) | | avg($\hat{\beta}_0$) | | avg($\hat{\beta}_1$) | | avg($\hat{\beta}_2$) | |
| DQ | – | – | – | – | – | – | −1.852 | −1.850 | 1.833 | 1.829 | −1.193 | −1.191 |
| SST | – | – | – | – | – | – | −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($\hat{\lambda}_1^T$) | | MSE($\hat{\lambda}_1^L$) | | MSE($\hat{\lambda}_0^T$) | | MSE($\hat{\beta}_0$) | | MSE($\hat{\beta}_1$) | | MSE($\hat{\beta}_2$) | |
| DQ | – | – | – | – | – | – | 0.827 | 0.771 | 0.189 | 0.116 | 0.094 | 0.053 |
| SST | – | – | – | – | – | – | 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($\hat{\lambda}_1^T$) | | avg($\hat{\lambda}_1^L$) | | avg($\hat{\lambda}_0^T$) | | avg($\hat{\beta}_0$) | | avg($\hat{\beta}_1$) | | avg($\hat{\beta}_2$) | |
| DQ | – | – | – | – | – | – | −1.323 | −1.325 | 1.931 | 1.928 | −1.256 | −1.251 |
| SST | – | – | – | – | – | – | −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($\hat{\lambda}_1^T$) | | MSE($\hat{\lambda}_1^L$) | | MSE($\hat{\lambda}_0^T$) | | MSE($\hat{\beta}_0$) | | MSE($\hat{\beta}_1$) | | MSE($\hat{\beta}_2$) | |
| DQ | – | – | – | – | – | – | 0.167 | 0.137 | 0.110 | 0.055 | 0.053 | 0.027 |
| SST | – | – | – | – | – | – | 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 |

*Notes*. DQ denotes estimated logistic regression parameters based on a complete case analysis of direct questioning responses only. SST denotes logistic regression parameters based on an analysis of sensitive question technique-based responses only (using an appropriately modified likelihood). Joint refers to logistic regression parameters based on the joint response model developed in the article. Two thousand random samples of size $n$ were drawn for all Monte Carlo experiments. For the SST and joint models, we set $p = 0.25$.

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, 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.

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 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 2 presents boxplots of the estimated APDs associated with setting $X_2$ to its 95th percentile value in a given sample instead of its 5th percentile

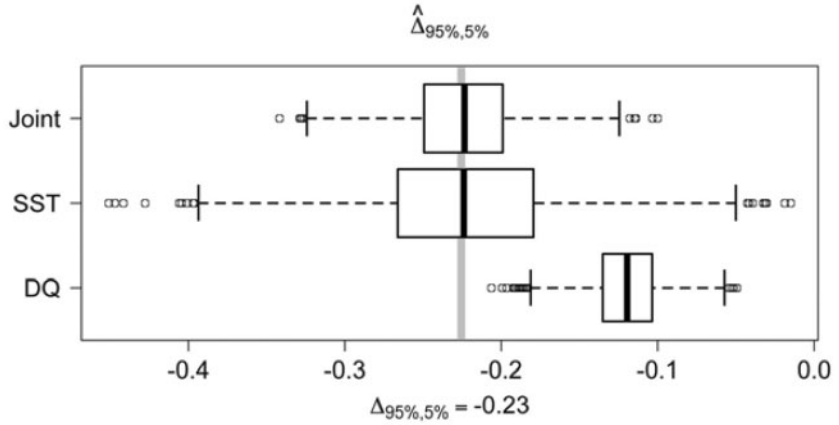


**Fig. 2** Boxplots of APDs across alternative estimators (Monte Carlo simulations).
*Notes:* The boxplots shown above are for Monte Carlo simulations for the high-evasiveness scenario with sample size $n = 5000$. The thick vertical bar denotes the true APD in the population.

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 Application to the Study of Corruption and Drug Use in Costa Rica

In this section of the article, we illustrate the utility of the joint response model developed in the article by using it to examine the results from a large-scale household survey about citizen experiences with crime and corruption conducted in Costa Rica. We do so in three steps. First, we use the data to illustrate the utility of the joint response model as a diagnostic tool for identifying the topics which do and do not require the protection afforded to respondents through the use of a SST. Second, we use the data to show how our model can be used to identify subgroups of the population for whom the protection afforded by a SST is especially important. Finally, we utilize the explanatory component of the joint response model in order to assess how demographic characteristics such as gender relate to an individual's willingness to bribe a police officer and the likelihood of having done so in the past.

The household survey was conducted in face-to-face format from October 2013 to April 2014. It was executed by *Borge y Asociados*, the most prominent survey research firm in Central America. The target population of the survey consisted of all residents 18 years or older residing in what is known as the *Gran Área Metropolitana* (GAM), which includes cantons in the provinces of San José, Heredia, Cartago, and Alajuela. The GAM is the principal urban center of Costa Rica. It contains approximately 2.6 million residents, and accounts for 60% of the country's entire population. The total sample size of the survey was equal to 4200 respondents.[3]

The instrument employed in the survey included a series of questions about sensitive issues. These included questions about corruption, drug use, and personal knowledge of individuals involved in the narcotics trade. The questions were presented twice to respondents. First, they were presented in the format of the crosswise model, in the manner shown in Fig. 1. Later, at

[3]The sampling strategy of the survey consisted of two-stage stratified random sampling (with fixed proportions defined for gender and age groupings).

the very end of the survey, the exact same questions were presented to respondents directly, with the explicit option of "prefer not to respond" given as a potential response. The innocuous group indicators utilized in the crosswise questions were based on mother's month of birth, father's month of birth (or that of another relative if father's birth month was unknown), and mother's day of birth. Specifically, the innocuous statement based on month of birth was "My mother was born in October, November, or December" or the equivalent for one's father. The innocuous statement based on day of birth was "My mother was born prior to the 10th day of the month (day 1 to 9)."

Survey enumerators went through extensive training in executing the crosswise component of the survey. This training consisted of a thorough explanation of the logic and functioning of the technique, as well as live practice sessions in which each enumerator practiced her delivery of this section of the survey with a compatriot in front of members of the research team and administrators from *Borge y Asociados*. According to the contract signed by the survey firm, only enumerators that had gone through the training sessions were permitted to deliver the surveys.[4]

In order to characterize the probabilities for the birthday-based group indicators, a nationally representative telephone survey of 1200 Costa Ricans was conducted in July 2013. The survey asked respondents directly about the month and day of birth of their mother and father. Since there should be no systematic differences in month and day of birth across sex of child, the probabilities were calculated by pooling the responses for mothers and fathers. To check the veracity of the survey reports, these were checked against statistical tables produced by Costa Rica's National Institute for Statistics and Censuses (INEC) on month of birth for newborns for the 2000–2011 period. According to the survey reports, the proportion of mother's and father's birthdays occurring in October, November, or December (the months employed in most of the crosswise questions) was 0.264. Averaging across the years 2000–2011, the actual proportion of newborn births falling into these months according to INEC was nearly identical, 0.265. Thus, there appears to be no evidence of recall bias or similar problems that might threaten the use of birthdays in the manner employed in this article.

### 6.1 *Variation in Evasiveness across Sensitive Topics*

The sensitive questions utilized in the survey differed along two key dimensions. As mentioned above, they differed in terms of topic, some dealing with bribery of police and others dealing with drug use and the drug trade. Some also differed in the degree to which the question prompted the respondent to directly implicate herself in having engaged in an illicit or socially undesirable act. In structuring the questions in this manner, our expectation was that evasiveness under direct questioning would likely be higher for the bribery questions than those about drug use, since bribery is unequivocally illegal in Costa Rica, and that, holding the topic constant, evasiveness under direct questioning would be higher the more the question was about the respondent's actual past behavior. Table 4 presents the wording for the sensitive questions, along with prevalence estimates from direct questioning, SST questioning only, and the full suite of parameter estimates provided by the baseline joint response model.

In both cases, our expectations were confirmed by the findings. The value of the parameter capturing the probability of a truthful response under direct questioning by those bearing the sensitive trait, $\hat{\lambda}_1^T$, was a good deal lower for questions dealing with willingness to bribe or a past history of bribery than it was for questions dealing with personal contact with drug users or a past history of drug use. The highest estimated level of evasiveness encountered across the questions was for question SQ2 in Table 5, which queried the respondent to indicate whether she

[4]An important feature of the delivery of this component of the survey consisted of a script describing to respondents how a hypothetical individual with a particular value on a sensitive item and a mother born in a particular month would respond to a given crosswise item. This script was delivered orally to all respondents prior to the commencement of the sensitive questions of interest. Focus group sessions conducted in San José with Costa Ricans of varied backgrounds in August 2013, prior to fielding the household survey, demonstrated that including a script of this sort considerably enhanced understanding and comfort with the technique.

**Table 4** Parameter estimates for five sensitive survey questions (calculated across estimation strategies)

| *SQ1: To avoid paying a traffic ticket, I would be willing to pay a bribe to a police officer* | | | | | |
|---|---|---|---|---|---|
| Prevalence estimate ($\hat{\pi}$) | | | Diagnostic parameters (joint model response model) | | |
| Direct only | 0.18 | [0.17,0.19] | $\hat{\lambda}_1^T$ | 0.61 | [0.55,0.69] |
| SST only | 0.22 | [0.18,0.25] | $\hat{\lambda}_1^L$ | 0.35 | [0.28,0.41] |
| Joint response | 0.29 | [0.26,0.32] | $\hat{\lambda}_0^T$ | 0.97 | [0.96,0.98] |
| Innocuous: My mother was born in October, November, or December ($p = 0.264$) | | | | | |
| SQ2: I have paid, at least once, a bribe to a police officer to avoid a traffic ticket | | | | | |
| Prevalence estimate ($\hat{\pi}$) | | | Diagnostic parameters (joint model response model) | | |
| Direct only | 0.09 | [0.08,0.10] | $\hat{\lambda}_1^T$ | 0.54 | [0.45,0.65] |
| SST only | 0.13 | [0.10,0.16] | $\hat{\lambda}_1^L$ | 0.44 | [0.32,0.53] |
| Joint response | 0.16 | [0.13,0.19] | $\hat{\lambda}_0^T$ | 0.98 | [0.97,0.99] |
| Innocuous: My mother was born in October, November, or December ($p = 0.264$) | | | | | |
| SQ3: Over the last year, I have had personal contact with a recreational drug user | | | | | |
| Prevalence estimate ($\hat{\pi}$) | | | Diagnostic parameters (joint model response model) | | |
| Direct only | 0.45 | [0.43,0.46] | $\hat{\lambda}_1^T$ | 0.90 | [0.86,0.94] |
| SST only | 0.37 | [0.34,0.40] | $\hat{\lambda}_1^L$ | 0.09 | [0.04,0.13] |
| Joint response | 0.49 | [0.46,0.51] | $\hat{\lambda}_0^T$ | 0.97 | [0.95,0.98] |
| Innocuous: My father was born in October, November, or December ($p = 0.264$) | | | | | |
| SQ4: Over the last year, I have consumed (at least once) some type of drug | | | | | |
| Prevalence estimate ($\hat{\pi}$) | | | Diagnostic parameters (joint model response model) | | |
| Direct only | 0.11 | [0.10,0.12] | $\hat{\lambda}_1^T$ | 0.78 | [0.64,0.96] |
| SST only | 0.10 | [0.07,0.13] | $\hat{\lambda}_1^L$ | 0.18 | [0.00,0.33] |
| Joint response | 0.14 | [0.11,0.16] | $\hat{\lambda}_0^T$ | 0.98 | [0.97,0.98] |
| Innocuous: My father was born in October, November, or December ($p = 0.264$) | | | | | |
| SQ5: I know personally someone involved in the sale, transport, or distribution of narcotics | | | | | |
| Prevalence estimate ($\hat{\pi}$) | | | Diagnostic parameters (joint model response model) | | |
| Direct only | 0.22 | [0.21,0.24] | $\hat{\lambda}_1^T$ | 0.72 | [0.65,0.80] |
| SST only | 0.22 | [0.19,0.25] | $\hat{\lambda}_1^L$ | 0.25 | [0.17,0.32] |
| Joint response | 0.30 | [0.27,0.33] | $\hat{\lambda}_0^T$ | 0.97 | [0.96,0.98] |
| Innocuous: My mother was born prior to the 10th day of the month (day 1 to 9) ($p = 0.278$) | | | | | |

*Note*. Ninety-five percent confidence intervals in square brackets.

had ever paid a bribe to a police officer to avoid a ticket. Respondents bearing the sensitive trait in this instance were only slightly more likely to tell the truth under direct questioning than they were to lie ($\hat{\lambda}_1^T = 0.54, \hat{\lambda}_1^L = 0.44$), indicating the value of the protection provided by the crosswise format. The lowest estimated level of evasiveness encountered across the questions was for question SQ3, which queried the respondent to indicate whether she had personal contact with a recreational drug user in the past year. For this question, the findings suggest that nearly all

respondents bearing the sensitive trait were willing to tell the truth under direct questioning ($\hat{\lambda}_1^T = 0.90, \hat{\lambda}_1^L = 0.09$), indicating that the use of an SST would not be advisable for that particular topic. Generally speaking, it appears that corruption is a topic for which the protection of an SST is required for this population, whereas this does not seem to be the case for questions about drug use.

Holding the topic constant, the estimated truthfulness parameters were also lower for questions prompting respondents to implicate themselves in past illicit behavior than for questions based on hypotheticals or contact with others engaged in illicit behavior. In this respect, respondents bearing the sensitive trait appeared more inclined to respond truthfully about willingness to bribe a police officer than about having actually done so (see SQ1 versus SQ2). The same appeared to be true for respondents bearing the sensitive trait when asked about contact with a recreational drug user as opposed to actual past personal drug use (SQ3 versus SQ4). Although consistent with expectations, it is important to treat these particular differences with caution, however, as they are relatively small in magnitude and not statistically significant by conventional standards.

### 6.2 *Variation in Evasiveness across Subgroups*

In the previous section, we provided evidence that the degree of sensitivity of a topic under study shapes the returns to using an SST. In other words, the variation in evasiveness across sensitive topics makes the use of a technique to deal with social desirability bias more or less crucial. In this section, we focus on how personal characteristics may affect respondents' willingness to provide a truthful response when asked directly. In particular, we examine the role of citizenship in shaping respondents' willingness to answer truthfully to direct questions about sensitive topics.

To explore variation across citizens and non-citizens, we use two different questions about corruption. The first one asks respondents about their willingness to pay a bribe to a police officer in order to avoid paying a traffic ticket (SQ1); the second one asks respondents whether they had actually bribed a police officer in the past to avoid a traffic ticket (SQ2). Our expectation is that non-citizens in Costa Rica will be less inclined to respond affirmatively to a hypothetical question about willingness to participate in an illegal activity and even less so when the question prompts respondents to implicate themselves in past illicit behavior. Given their precarious status, non-citizens may be more wary of the potential legal consequences of their answers and, as a consequence, they may be more reluctant to respond honestly when asked directly about an illegal activity such as paying a bribe. Figure 3 presents the prevalence estimates from direct questioning, SST questioning only, and the parameter estimates provided by the joint response model across Costa Ricans and Non-Costa Ricans.[5]

Figure 3 shows that, in line with our expectations, Costa Ricans and Non-Costa Ricans respond differently when asked about an illegal behavior such as corruption. First, as in the previous section, the estimated truthfulness parameters for both Costa Ricans and Non-Costa Ricans were lower for the question about previous illicit behavior than for the question about hypothetical willingness to engage in this behavior. Second, the value of the parameter that captures the probability of a truthful response under direct questioning by those who are actually willing to bribe or had bribed in the past, $\hat{\lambda}_1^T$, is lower for Non-Costa Ricans than it is for Costa Ricans. When asked directly about willingness to bribe a police officer, both groups report an affirmative response proportion of 0.18. However, both the SST and the joint response model—with the protection provided by these techniques—show a considerably higher affirmative response for Non-Costa Ricans (0.31 and 0.37, respectively), while the difference for Costa Ricans is noticeably smaller (0.21 and 0.28). Indeed, in the case of Non-Costa Ricans, those bearing the sensitive trait were about equally likely to tell the truth and lie under direct questioning ($\hat{\lambda}_1^T = 0.45, \hat{\lambda}_1^L = 0.48$), whereas

[5]Recall that the survey was a household survey, which means that the Non-Costa Ricans included in the survey are residents, not tourists. Most non-nationals living in Costa Rica come from Nicaragua, having migrated to Costa Rica (either permanently or temporarily) primarily for economic reasons. Of the 4200 respondents of the survey, 394 (9.4%) were Non-Costa Ricans; of these, 344 (87.3%) were from Nicaragua.

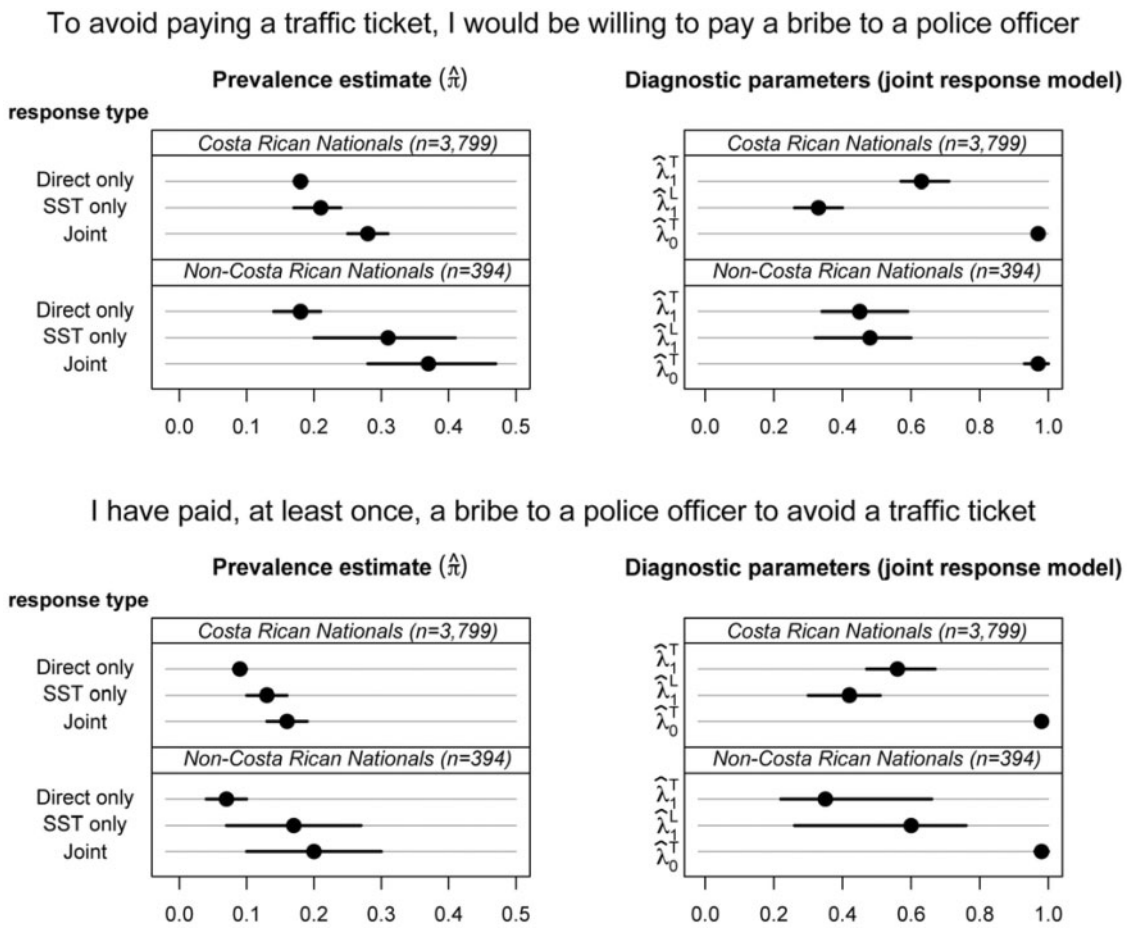


**Fig. 3** Parameter estimates for questions on corruption (broken down by citizenship status). Notes: Point estimates for model parameters denoted by large black circles. Ninety-five percent confidence intervals denoted by horizontal black lines.

for Costa Ricans the findings show that many more respondents bearing the sensitive trait were willing to tell the truth under direct questioning ($\hat{\lambda}_1^T = 0.63$, $\hat{\lambda}_1^L = 0.33$).

For the question that asked respondents to directly implicate themselves in having engaged in illegal behavior in the past, we also found that the probability of a truthful response under direct questioning by those who had bribed in the past is substantially lower for Non-Costa Ricans ($\hat{\lambda}_1^T = 0.35$) than it is for Costa Ricans ($\hat{\lambda}_1^T = 0.56$).

Taken together, these results show that the potential variation across subgroups might make it necessary to use an SST, even when the topic does not seem to be particularly sensitive for the population as a whole. In particular, vulnerable groups like immigrants appear more reluctant to provide a truthful response when asked directly about illegal behavior.

### 6.3 *Gender Differences Toward Bribery of the Police*

The literature on corruption has grown significantly in the past decade, offering a wealth of evidence regarding the individual and macro determinants of corrupt behavior. One branch of the literature on the individual determinants of corruption has concentrated on elucidating gender differences. Although understanding of the theoretical mechanisms at play is still incomplete, a body of empirical work suggests that women are generally less involved in corruption than men and also less likely to condone corruption. Indeed, it has been argued that countries that have larger representation of women in Parliament, in senior positions in the government bureaucracy, and in private-sector management have lower levels of perceived corruption (Dollar, Fisman, and

Gatti 2001; Swamy et al. 2001).[6] The explanation for this cross-country finding is based on the fact that women, when surveyed, tend to report lower levels of tolerance for corruption. For instance, using data from the World Values Survey, Swamy et al. (2001) show that women are less likely to condone bribe taking. In the same paper, but using survey data from a World Bank Study on Georgia, these authors also show that officials in firms owned or managed by men are significantly more likely to be involved in bribe-giving than those managed or owned by women. Similarly, using data on eight Western European countries from the World Values Survey and the European Values Survey, Torgler and Valev (2010) show the existence of significantly greater aversion to corruption and tax evasion among women. More recently, it has been argued that the reason women may be less likely to engage in and tolerate corruption is because women are more risk averse and therefore more inclined to follow norms (Esarey and Chirillo 2013).

In this final section of the article, we utilize the explanatory joint response model developed in the text in order to evaluate the aforementioned literature's expectations about the relationship between gender and predilections toward corruption. Our outcomes are both attitudinal and behavioral. The attitudinal outcome consists of a respondent's joint response about her willingness to bribe a police officer in order to avoid a traffic ticket. The behavioral outcome consists of a respondent's response about whether she had at any point in the past paid a bribe to a police officer to avoid a ticket. In addition to gender, we include in our estimations additional demographic characteristics such as the logarithm of the respondent's age, her level of education, indicators of socioeconomic status such as possession of a car, laptop, tablet, and internet in the home, as well as indicators that tap into the nature of the respondent's social networks, specifically, whether she knows personally a police officer or knows personally someone accused, prosecuted, or sentenced in the criminal justice system. In all estimations, we utilized an inverse logit link function for the explanatory component of the joint response model.

Table 5 displays the parameter estimates from the two estimated models. With respect to gender, both models tell a similar story: men are substantially more likely than women to indicate a willingness to bribe as well as to indicate that they have done so in the past. The APDs provided in the table give a sense of the strength of the conditional association between gender and bribery. These numbers can be interpreted as follows. Based on the estimates in the first model, if one were to construct two samples of the same size as the full sample, one consisting solely of men and another consisting solely of women, with the individuals in the two samples sharing the exact same background characteristics save for gender (i.e., those encountered among the actual respondents in the full sample), the expected proportion of individuals willing to bribe in the all-male sample would be 0.14 greater than the expected proportion in the all-female sample. Doing the same and using the estimates of the second model, one would find that the expected proportion of individuals with an actual history of bribery in the all-male sample would be 0.13 higher than the expected proportion in the all-female sample. In both cases, the APDs are statistically significant by any conventional standard. Thus, our analysis of the data from Costa Rica clearly affirms the expectations of the growing literature on gender differences in proclivities toward corruption.[7]

In terms of the other background characteristics, our analysis finds that younger respondents were substantially more inclined to indicate a willingness to bribe than older respondents (but were no more likely to have had a history of bribery), respondents with low to intermediate levels of education were more likely than college-educated respondents to indicate both a willingness to bribe and a history of doing so, respondents belonging to a social network including a police officer were—disturbingly—*more* inclined to indicate a willingness to bribe than those not belonging to such a network, and respondents belonging to a social network including someone implicated in criminal activity were

[6]However, Sung (2003) and Goetz (2007) argue that the correlation between the participation of women in politics and lower levels of corruption might be spurious and caused by other aspects of liberal democracy that tend to go together with gender equality.

[7]Importantly, our findings in this section do not appear to be a function of differences between men and women in the willingness to respond truthfully under direct questioning. Prior to conducting our analysis, we divided our sample according to gender and estimated all relevant diagnostic parameters separately. For both questions, the diagnostic parameters were quite similar for men and women, thereby justifying a pooled analysis as conducted here. Full results are available upon request.

**Table 5** Relationship between gender and bribery of police (joint response model)

| | *Outcome: Willing to pay a bribe* | | | *Outcome: Paid a bribe* | | |
|---|---|---|---|---|---|---|
| Parameters | Estimate | s.e. | 95% int. | Estimate | s.e. | 95% int. |
| Diagnostic parameters | | | | | | |
| $\hat{\lambda}_1^T$ | 0.66 | 0.03 | [0.60,0.72] | 0.63 | 0.06 | [0.54,0.76] |
| $\hat{\lambda}_1^L$ | 0.31 | 0.03 | [0.25,0.36] | 0.35 | 0.06 | [0.23,0.44] |
| $\hat{\lambda}_0^T$ | 0.97 | 0.00 | [0.96,0.98] | 0.98 | 0.00 | [0.97,0.99] |
| Explanatory parameters | | | | | | |
| Constant | 2.41 | 0.50 | [1.51,3.46] | −3.71 | 0.62 | [−4.90, −2.53] |
| **Male** | **0.79** | **0.10** | **[0.60,1.00]** | **1.23** | **0.13** | **[0.99,1.50]** |
| log(age) | −1.34 | 0.13 | [−1.61, −1.07] | −0.01 | 0.15 | [−0.28,0.28] |
| Education (base = some college) | | | | | | |
| Primary or less | 0.33 | 0.17 | [0.02,0.67] | 0.11 | 0.24 | [−0.34,0.58] |
| Secondary incomplete | 0.54 | 0.16 | [0.23,0.89] | 0.19 | 0.19 | [−0.16,0.57] |
| Secondary complete | 0.34 | 0.16 | [0.08,0.68] | 0.45 | 0.20 | [0.09,0.89] |
| Some technical | 0.08 | 0.29 | [−0.48,0.60] | 0.34 | 0.31 | [−0.32,0.97] |
| Laptop | 0.05 | 0.12 | [−0.20,0.28] | 0.14 | 0.15 | [−0.16,0.42] |
| Tablet | 0.12 | 0.12 | [−0.10,0.38] | 0.41 | 0.14 | [0.13,0.67] |
| Car | 0.24 | 0.12 | [0.04,0.47] | 0.48 | 0.13 | [0.23,0.75] |
| Internet | 0.03 | 0.13 | [−0.24,0.26] | 0.02 | 0.16 | [−0.28,0.33] |
| Knows a police officer | 0.19 | 0.10 | [0.03,0.39] | 0.20 | 0.13 | [−0.05,0.45] |
| Knows someone accused, prosecuted, or sentenced by the justice system | 0.61 | 0.10 | [0.39,0.80] | 0.74 | 0.13 | [0.51,1.03] |
| **APD (male versus female)** | **0.14** | **0.02** | **[0.10,0.18]** | **0.13** | **0.02** | **[0.10,0.16]** |
| | $n = 4072$ | | | $n = 4066$ | | |

Bold values indicate the substantive focus of the subsection.

substantially more inclined to indicate both a willingness to bribe and a history of having done so than those not belonging to such a network. As is to be expected given the emphasis in our questions on bribery as a means of avoiding fines for traffic infractions, ownership of a car was associated with both a greater willingness to bribe and a greater likelihood of having done so.

## 7 Conclusion

This article 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 on SSTs as well as on 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 an SST to study the sensitive behavior of interest; (2) efficiently estimate the prevalence of the sensitive behavior; and (3) efficiently 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 our approach follows directly from an assumption we dubbed 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 on responses generated by SSTs do *not* do so.

This is an important advantage of our approach, 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, we recognize that this article nevertheless leaves much to be done. An implicit assumption of our 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 direct questioning 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 our approach.

A broader question that is raised by our article, one that needs to be addressed in any application that asks respondents about sensitive behaviors using both an SST and direct questioning, concerns the responsibility of survey researchers to protect respondents from harm. Even if some respondents may be willing to directly reveal their having engaged in a sensitive behavior, there are settings in which we would deem it inappropriate for a survey to prompt them to do so. In particular, situations in which direct reporting of a sensitive behavior might expose respondents to physical coercion are ones in which we would be disinclined to pursue a questioning strategy that prompts respondents to directly reveal participation in a sensitive behavior. As a practical matter, this means that we would not recommend that our framework be applied in areas undergoing active military conflict or to criminal activities carrying with them a non-negligible probability of violent reprisal. In such circumstances, we recognize that solely the use of an SST making it impossible to link responses to possession of the sensitive behavior is likely to be the best option. However, for a wide variety of behaviors typically studied today using only an SST or direct questioning, such as low-level corruption or fraud, vote buying, or forms of racial or ethnic bias, the joint response approach developed here would be both appropriate and may offer substantial advantages over existing approaches.

## Funding

The research for this paper was conducted with generous support from the Inter-American Development Bank.

*Conflict of interest statement*. None declared.

## References

Aronow, P. M., A. Coppock, F. W. Crawford, and D. P. Green. 2015. Combining list experiment and direct question estimates of sensitive behavior prevalence. *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smu023.

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. 2014. Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science* 58(4):1043–63.

Blair, G., K. Imai, and Y.-Y. Zhou. 2015. Statistical analysis of the randomized response technique. *Journal of the American Statistical Association* 110(511):1304–19.

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):245–62.

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* 24(3):377–92.

Bourke, P. D., and M. A. Moran. 1988. Estimating proportions from randomized response data using the EM algorithm. *Journal of the American Statistical Association* 83(404):964–68.

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):543–64.

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* 39(1):1–38.

Dollar, D., R. Fisman, and R. Gatti. 2001. Are women really the "fairer" sex? Corruption and women in government. *Journal of Economic Behavior & Organization* 46(4):423–29.

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.

Esarey, Justin, and Gina Chirillo. 2013. "Fairer sex" or purity myth? Corruption, gender, and institutional context. *Politics and Gender* 9(4):390–413.

Fox, J. P. 2005. Randomized item response theory models. *Journal of Educational and Behavioral statistics* 30(2):189–212.

Fox, J. P., and C. Wyrick. 2008. A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics* 33(4):389–415.

Fox, J. P., 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.

Fox, J. P., M. Avetisyan, and J. Palen. 2013. Mixture randomized item-response modeling: A smoking behavior validation study. *Statistics in Medicine* 32(27):4821–37.

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

Gilens, M., P. M. Sniderman, and J. H. Kuklinski. 1998. Affirmative action and the politics of realignment. *British Journal of Political Science* 28(1):159–83.

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* 18:349–80.

Gingerich, D. W., V. Oliveros, A. Corbacho, and M. Ruiz-Vega. 2015. Replication files for "When to protect? Using the crosswise model to integrate protected and direct responses in surveys of sensitive behavior." Available on the Political Analysis Dataverse of Harvard University's Dataverse Network. http://dx.doi.org/10.7910/DVN/2AIHQF.

Glynn, A. N. 2013. What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly* 77(S1):159–72.

Goetz, A. M. 2007. Political cleaners: Women as the new anti-corruption force? *Development and Change* 38:87–105.

Gonzalez-Ocantos, E., C. K. De Jonge, C. Meléndez, J. Osorio, and D. W. Nickerson. 2012. Vote buying and social desirability bias: Experimental evidence from Nicaragua. *American Journal of Political Science* 56(1):202–17.

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.

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):1387–403.

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):402–19.

Lamb C. W. Jr. and D. E. Stem Jr. 1978. An empirical validation of the randomized response technique. *Journal of Marketing Research* 15(4):616–21.

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* 94:741–52.

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. D. Gueorguiev, and N. M. Jensen. 2015. Monopoly money: Foreign investment and bribery in Vietnam, a survey experiment. *American Journal of Political Science* 59(2):419–39.

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

Rosenfeld, B., K. Imai, and J. Shapiro. 2014. An empirical validation study of popular survey methodologies for sensitive questions. Unpublished manuscript, Princeton University.

Sung, H.-E. 2003. Fairer sex or fairer system? Gender and corruption revisited. *Social Forces* 82(2):703–23.

Swamy, Anand, Stephen Knack, Young Lee, and Omar Azfar. 2001. Gender and corruption. *Journal of Development Economics* 64:25–55.

Tan, M. T., G. L. Tian, and M. L. Tang. 2009. Sample surveys with sensitive questions: A nonrandomized response approach. *American Statistician* 63(1):9–16.

Torgler, B., and N. T. Valev. 2010. Gender and public attitudes toward corruption and tax evasion. *Contemporary Economic Policy* 28(4):554–68.

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):251–63.