List Experiments with Measurement Error*

Graeme Blair† Winston Chou‡ Kosuke Imai§

April 21, 2018

Abstract

Measurement error threatens the validity of survey research especially when studying sensitive questions. In the context of list experiments, Ahlquist (2017) introduces the notion of "top-biased" response error, in which a random fraction of respondents provide the maximal response regardless of their truthful answer to the sensitive question. Ahlquist conducts simulation studies based on this scenario and finds that the maximum likelihood (ML) regression estimator, proposed in Imai (2011) and further extended in Blair and Imai (2012), exhibits severe model misspecification bias when the sensitive trait is rare. Unfortunately, Ahlquist stops short of offering any solution to the general problem of measurement error in list experiments. We take up this challenge and provide new tools for diagnosing and mitigating measurement error in list experiments. First, we point out that top-biased error is unlikely for truly sensitive questions, as it implies that respondents are willing to admit having a sensitive trait even when they do not. Second, we show that the nonlinear least squares (NLS) regression estimator is robust to top-biased error. Third, we consider an alternative form of response error, mentioned but not studied in Ahlquist (2017), in which a small fraction of respondents offer a random response to the list experiment. We show that both ML and NLS regression estimators are robust to such error. Fourth, we propose a statistical test for detecting general model misspecification caused by misreporting. Fifth, we demonstrate how to directly model nonstrategic respondent error and how to build a more robust regression model. Finally, we reanalyze the empirical examples studied in Ahlquist (2017) and demonstrate that simple diagnostic tools can be used to avoid the problems identified in the original article. We conclude this article with a set of practical recommendations for applied researchers. The proposed methods are implemented through an open-source software package.

Key Words: auxiliary information, indirect questioning, item count technique, misspecification test, sensitive survey questions, unmatched count technique

*All the proposed methods presented in this paper are implemented as part of the R package, `list: Statistical Methods for the Item Count Technique and List Experiment`, which is freely available for download at http://cran.r-project.org/package=list (Blair, Chou and Imai, 2017).

†Assistant Professor of Political Science, UCLA. Email: graeme.blair@ucla.edu, URL: `https://graemeblair.com`

‡Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: wchou@princeton.edu, URL:`http://princeton.edu/~wchou`

§Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: `https://imai.princeton.edu`

1 Introduction

Measurement error threatens the validity of survey research. This is especially so when studying sensitive topics due to the incentives for respondents to misrepresent the truth. Along with other methods such as the randomized response and endorsement experiments (Blair, Imai and Zhou, 2015; Bullock, Imai and Shapiro, 2011; Gingerich, 2010), the list experiment (a.k.a. the item count technique and the unmatched count technique) is an indirect questioning method that seeks to mitigate potential biases from social desirability and nonresponse by veiling individual responses (Corstange, 2009; Imai, 2011; Blair and Imai, 2012; Glynn, 2013). While some studies have shown that list experiments can be effective for reducing bias, a well-known limitation is that the extreme value responses perfectly reveal the sensitive trait, making some respondents disguise their truthful answers. Blair and Imai (2012) show how to address such strategic measurement error by modeling floor and ceiling effects within a regression framework.

Ahlquist (2017) criticizes the methodological literature on list experiments for "ignoring the implications of arguably more common nonstrategic measurement error due to the usual problems of miscoding by administrators or enumerators as well as respondents misunderstanding or rushing through surveys" (p. 8). The paper introduces the *top-biased error* mechanism, in which a small, random fraction of respondents provide the maximal response value regardless of their truthful answer to the sensitive question. Based on simulation studies, the paper finds that the maximum likelihood regression estimator (MLreg), proposed in Imai (2011) and extended in Blair and Imai (2012), suffers from a greater degree of bias than the Difference-in-Means estimator (DiM) when the sensitive trait is rare. Unfortunately, Ahlquist (2017) stops short of offering any general solution to the problems that the article identifies, stating "We do not yet have tools for determining the levels, rates, and structure of nonstrategic respondent error and developing such tools seems unlikely" (p. 16).

In this paper, we take up the challenge of nonstrategic measurement error in list experiments and develop new statistical methods for detecting measurement error and alleviating the resulting model misspecification bias. As a preliminary observation, we point out that the comparison between MLreg and DiM, as drawn in Ahlquist (2017), is inappropriate because the two estimators are aimed at different goals. Whereas DiM is concerned with the prevalence of a sensitive trait, MLreg sheds light on the multivariate relationship between respondent covariates and the sensitive trait, and so requires additional assumptions. Therefore, we propose to compare MLreg with the nonlinear least squares regression estimator (NLSreg), originally proposed in Imai (2011) as a generalization of DiM, rather than with DiM. We provide a formal means of conducting this comparison and show that NLSreg retains the robustness of DiM while still empowering researchers

to study multivariate relationships.

Importantly, we also offer tools for addressing multiple measurement error mechanisms. This is because, as we argue in Section 2.2, the top-biased error mechanism is implausible for truly sensitive questions, since the maximal response reveals that respondents have the sensitive trait. Thus, top-biased error runs directly counter to respondents' incentives to misrepresent the truth. A more plausible nonstrategic measurement error mechanism, discussed though not studied in detail in Ahlquist (2017), is *uniform error*, in which a small fraction of respondents provide a random response. We find in our simulation study that MLreg is reasonably robust to such error.

In our view, all forms of nonstrategic measurement error are best avoided through careful enumerator training, pilot surveys, and other best practices of survey research. Nevertheless, we show how to detect and adjust for top-biased and uniform error in a regression modeling framework (Section 2). First, as mentioned earlier, we show that NLSreg, also proposed in Imai (2011), is more robust to model misspecification than MLreg. Second, we develop a statistical test to detect the presence of measurement error, based on the idea that the difference between NLSreg and MLreg implies empirical evidence for model misspecification. Third, following Blair and Imai (2012), we demonstrate how to model nonstrategic measurement error mechanisms using NLSreg and MLreg. These regression models contain the model without measurement error as a limiting case, providing another statistical test. Fourth, we propose a method for improving the robustness of the standard regression estimators using the auxiliary information strategy of Chou, Imai and Rosenfeld (2017). All of our proposed methods are implemented via the open-source R package `list` (Blair, Chou and Imai, 2017).

We demonstrate the performance of the proposed methodology by revisiting the simulation studies in Ahlquist (2017) (Section 3). We show that our proposed model misspecification test detects deviations from the modeling assumptions at a high rate. We also confirm the theoretical expectation that NLSreg is robust to nonstrategic measurement error and the forms of model misspecification contemplated in Ahlquist (2017). Turning to uniform response error, we find that MLreg performs reasonably well despite this misspecification. Nevertheless, we show that the robust estimators proposed here and in Chou, Imai and Rosenfeld (2017) can improve the performance of list experiment regression in the presence of both types of measurement error.

Finally, we apply the proposed methodology to the empirical study presented in Ahlquist (2017) (Section 4). The study claims to show that, when analyzed via MLreg, unrealistically large proportions of Americans engage in voter fraud and/or were abducted by aliens. By contrast, a list experiment on texting while driving did not reveal such problems. The most straightforward analysis of these data via DiM yields a negative estimate (a positive but statistically insignificant estimate)

for the proportion of those who engage in voter fraud (were abducted by aliens). We caution that multivariate analysis of such list experiments is bound to be unreliable, as a trait needs to exist for it to covary with respondent characteristics. Nevertheless, we show that our methods yield more sensible estimates of the prevalence of these traits than MLreg. In particular, our uniform error model yields estimates of voter fraud and alien abduction that are statistically indistinguishable from zero with the narrowest confidence interval among the estimators we consider.

We further demonstrate that for the list experiment on texting while driving, which is not an extremely rare event, MLreg yields reasonable results that agree with those of the other methods. Given that all three list experiments were conducted by the same researchers on the same respondents, and so were likely subject to the same forms of measurement error, this finding indicates that researchers should chiefly be concerned with the rarity of the sensitive traits when deciding whether multivariate regression analyses are appropriate. Building on this observation, we conclude this article by offering a set of practical recommendations for applied researchers conducting list experiments (Section 5).

2 The Proposed Methodology

In this section, we propose statistical methods for analyzing list experiments with measurement error. We begin by reviewing MLreg and NLSreg, introduced in Imai (2011) and extended in Blair and Imai (2012). We then propose a statistical test of model misspecification for detecting measurement error. Next, following Blair and Imai (2012), we show how to directly model measurement error mechanisms and apply this strategy to the top-biased and uniform error processes introduced in Ahlquist (2017). Finally, we adopt another modeling strategy developed in Chou, Imai and Rosenfeld (2017) to further improve the robustness of multivariate regression models.

2.1 Multivariate Regression Models: A Review

Suppose that we have a simple random sample of N respondents from a population. In standard list experiments, we have a total of J binary control questions and one binary sensitive question. Let T_i be the randomized treatment assignment indicator. That is, $T_i = 1$ indicates that respondent i is assigned to the treatment group and is asked to report the total number of affirmative responses to the $J + 1$ items (J control items plus one sensitive item). In contrast, $T_i = 0$ implies that the respondent is assigned to the control group and is asked to report the total number of affirmative answers to J control questions. We use X_i to represent the set of K pre-treatment covariates.

Let Y_i denote the observed response. If respondent i belongs to the treatment group, this variable can take any non-negative integer less than or equal to $J + 1$, i.e., $Y_i \in \{0, 1, \ldots, J + 1\}$. On the other hand, if the respondent is assigned to the control group, the maximal value is J,

i.e., $Y_i \in \{0, 1, \ldots, J\}$. Furthermore, let Z_i represent the latent binary variable indicating the affirmative answer to the sensitive question. If we use the Y_i^* to represent the total number of affirmative answers to the J control questions, the observed response can be written as,

$$Y_i \;=\; T_i Z_i + Y_i^*. \tag{1}$$

In the early literature on list experiments, researchers estimated the proportion of respondents with the affirmative answer to the sensitive item using DiM, but could not characterize the respondents most likely to have the affirmative response. To overcome this challenge, Imai (2011) considers the following multivariate regression model,

$$\mathbb{E}(Y_i \mid T_i, X_i) \;=\; T_i \mathbb{E}(Z_i \mid X_i) + \mathbb{E}(Y_i^* \mid X_i) \tag{2}$$

where the randomization of treatment assignment guarantees the following statistical independence relationships, $T_i \perp\!\!\!\perp Z_i \mid X_i$ and $T_i \perp\!\!\!\perp Y_i^* \mid X_i$.

Although this formulation can accommodate various regression models, one simple parametric model, considered in Imai (2011), is the following binomial logistic regression model,

$$Z_i \mid X_i \;\overset{\text{indep.}}{\sim}\; \text{Binom}(1, g(X_i; \beta)) \tag{3}$$

$$Y_i^* \mid X_i \;\overset{\text{indep.}}{\sim}\; \text{Binom}(J, f(X_i; \gamma)) \tag{4}$$

where $f(X_i; \gamma) = \text{logit}^{-1}(X_i^\top \gamma)$ and $g(X_i; \beta) \;=\; \text{logit}^{-1}(X_i^\top \beta)$, implying the following regression functions, $\mathbb{E}(Y_i^* \mid X_i) = J \cdot f(X_i; \gamma)$ and $\mathbb{E}(Z_i \mid X_i) = g(X_i; \beta)$.

Imai (2011) proposes two ways to estimate this multivariate regression model: nonlinear least squares (NLSreg) and maximum likelihood (MLreg) estimation. NLSreg is obtained by minimizing the sum of squared residuals based on equation (2).

$$(\hat{\beta}_{\mathsf{NLS}}, \hat{\gamma}_{\mathsf{NLS}}) \;=\; \arg\min_{(\beta,\gamma)} \sum_{i=1}^{N} \{Y_i - T_i \cdot g(X_i; \beta) - f(X_i; \gamma)\}^2 \tag{5}$$

NLSreg is consistent so long as the regression functions are correctly specified and does not require the distributions to be binomial. One can obtain more efficient estimates by relying on distributional assumptions. In particular, MLreg is obtained by maximizing the following log-likelihood function,

$$\begin{aligned}
(\hat{\beta}_{\mathsf{ML}}, \hat{\gamma}_{\mathsf{ML}}) \;=\; & \underset{(\beta,\gamma)}{\text{argmax}} \sum_{i \in \mathcal{J}(1,0)} [\log\{1 - g(X_i; \beta)\} + J \cdot \log\{1 - f(X_i; \gamma)\}] \\
& + \sum_{y=0}^{J} \sum_{i \in \mathcal{J}(0,y)} y \log f(X_i; \gamma) + (J - y) \log\{1 - f(X_i; \gamma)\} + \sum_{i \in \mathcal{J}(1,J+1)} \{\log g(X_i; \beta) + J \log f(X_i; \gamma)\} \\
& + \sum_{y=1}^{J} \sum_{i \in \mathcal{J}(1,y)} \log \left[g(X_i; \beta) \binom{J}{y-1} f(X_i; \gamma)^{y-1} \{1 - f(X_i; \gamma)\}^{J-y+1} + \right.
\end{aligned}$$

$$\{1 - g(X_i; \beta)\} \binom{J}{y} f(X_i; \gamma)^y \{1 - f(X_i; \gamma)\}^{J-y} \Bigg] \tag{6}$$

where $\mathcal{J}(t, y)$ represents the set of respondents who have $T_i = t$ and $Y_i = y$.

The choice between NLSreg and MLreg involves a fundamental tradeoff between bias and variance. MLreg is more efficient than NLSreg because the former makes an additional distributional assumption. In particular, as Ahlquist (2017) points out, MLreg models each cell of the observed response, including the $Y_i = J+1$ and $Y_i = 0$ cells in the treatment group, which can greatly affect the parameter estimates. In contrast, NLSreg only makes an assumption about the conditional mean functions and hence is more robust to measurement errors in specific cells. Therefore, it is no surprise that Ahlquist (2017) finds DiM, which is a special case of NLSreg without covariates, is more robust for estimating the proportion of sensitive trait than MLreg. However, this comparison is inappropriate. The goal of NLSreg and MLreg is multivariate regression analysis. If one wishes merely to estimate the proportion of sensitive trait, DiM is sufficient and fitting a regression model that requires additional functional-form and/or distributional assumptions is unnecessary.

Two identification assumptions are required for DiM, NLSreg, and MLreg. First, respondents in the treatment group are assumed not to lie about the sensitive item, i.e., no liars. Any other behavior implies misreporting, and any estimator based on mismeasured responses is likely to be biased. The second assumption is that respondents' answers to the control items are not affected by the treatment, i.e., no design effect. Because list experiments rely upon the comparison of responses between the treatment and control groups, responses to the control items must remain identical in expectation between the two groups. The violation of this assumption also leads to mismeasured responses, yielding biased estimates. We emphasize that DiM is a special case of NLSreg and is not exempt from these assumptions. NLSreg adds an assumption about the correctly specified regression function and MLreg imposes an additional distributional assumption. In no way does DiM "not require the no liars assumption" (p. 15) as Ahlquist (2017) claims (In Appendix A, we prove that DiM is biased under the two nonstrategic error processes proposed by Ahlquist (2017). The bias is large when the prevalence of sensitive trait is small.).

The main difficulty of multivariate regression analysis for list experiments stems from the fact that the response to the sensitive item is not observed except for the respondents in the treatment group who choose the maximal or minimal response. There are two implications. First, regression analysis under this circumstance is more challenging than the usual situation, where the outcome is directly observed. If the sensitive trait of interest is a rare event, then MLreg is likely to suffer from bias. Such bias is known to exist even when the outcome variable is observed (King and Zeng, 2001), and is likely to be amplified for list experiments as demonstrated by Ahlquist (2017). Second, dealing with measurement error will be more difficult when the outcome variable is not

directly observed. Below, we consider several methodological strategies for addressing this issue.

2.2 Strategic and Nonstrategic Measurement Errors

Ahlquist (2017) introduces a nonstrategic measurement error mechanism, called *top-biased error*, in which a random fraction of respondents are assumed to give the maximal response value regardless of their truthful answer to the sensitive question. The motivation for this measurement error mechanism is unclear, although miscoding by survey administrators and respondents' confusion and careless errors are listed as possible reasons. Indeed, the article states (p. 5):

> I emphasize top-biased error not because there is any reason to believe that it is prevalent in applied situations but rather because top-biased error is likely to be the most problematic for both the DiM and ICT-ML [ML multivariate regression] estimators.

In fact, we do not believe that top-biased error is common in practice. The reason is simple. Under the treatment condition, giving the maximal value reveals that the respondent answers the sensitive question affirmatively. This implies, for example, that respondents are willing to admit engaging in such sensitive behaviors as drug use and tax evasion or having socially undesirable attitudes such as gender and racial prejudice. This scenario is highly unlikely so long as the behavior or attitudes researchers are trying to measure are actually sensitive.

For this reason, researchers are often rightly concerned with strategic measurement error, which originates from the incentives of respondents' to conceal their response to the sensitive item. Indeed, this is why researchers consider the use of indirect questioning techniques in the first place. In the context of list experiments, these *strategic* measurement errors result in floor and ceiling effects, in which some respondents avoid reporting (rather than gravitate to as assumed under the top-biased error) the extreme values.

As an example of strategic measurement error, we present a list experiment conducted by Lyall, Blair and Imai (2013) in the violent heart of Taliban-controlled Afghanistan, which was designed for estimating the level of support for the Taliban. The control group was given the following script ($J = 3$),

> `I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.`
>
> `Karzai Government; National Solidarity Program; Local Farmers`

For the treatment group, the sensitive actor, i.e., the Taliban, is added.

	Control group		Treatment group	
response	counts	percentage	counts	percentage
0	188	20	0	0
1	265	29	433	47
2	265	29	287	31
3	200	22	198	22
4			0	0

Table 1: An Example of Floor and Ceiling Effects from the List Experiment in Afghanistan Reported in Lyall, Blair and Imai (2013). No respondent in the treatment group gave an answer of 0 or 4, suggesting that the respondents were avoiding to reveal whether they support the Taliban.

```
Karzai Government; National Solidarity Program; Local Farmers; Taliban
```

Table 1 presents the descriptive information, which shows clear evidence of floor and ceiling effects. Indeed, no respondent gave an answer of 0 or 4. By avoiding the extreme responses of 0 and 4, respondents in the treatment group are able to remain ambiguous as to whether they support or oppose the Taliban. This strategic measurement error may have arisen in part because of the public nature of interview. As explained in Lyall, Blair and Imai (2013), interviewers are required to ask survey questions to respondents in public while village elders watch and listen. Under this circumstance, it is no surprise that respondents try to conceal their truthful answers. Because of this sensitivity, the authors of the original study used endorsement experiments (Bullock, Imai and Shapiro, 2011), which represent a more indirect questioning technique, in order to measure the level of support for the Taliban. On the other hand, Blair, Imai and Lyall (2014) find that in the same survey the list experiment works well for measuring the level of support for the international forces, which are a less sensitive actor to admit support or lack thereof for than are the Taliban.

In sum, we believe that in list experiments gauging truly sensitive issues, researchers should be concerned more about strategic measurement error than nonstrategic measurement error. Because list experiments deal with more sensitive items than standard survey questions, respondents have a greater incentive to conceal truthful answers to the sensitive question, thus encouraging floor and ceiling effects. In contrast, top-biased error assumes that respondents are willing to admit their sensitive traits, which goes directly against this incentive.

2.3 Detecting Measurement Error

Although researchers are unlikely to know the magnitude of measurement error, whether strategic or not, we can sometimes detect the measurement error from data. In addition to the tests developed by Blair and Imai (2012) and Aronow et al. (2015), we extend and formalize the recommendation by Ahlquist (2017) to compare DiM and MLreg. As mentioned earlier, this comparison is problematic because the two estimators are aimed at different goals. DiM is designed to estimate the prevalence of sensitive trait whereas MLreg is used to analyze multivariate relationships between the sensitive

trait and respondents' characteristics. Moreover, in Ahlquist (2017) there is no discussion of how to formally assess the similarity or difference of these estimates, making it difficult for applied researchers to determine how large the difference has to be in order to abandon the results based on MLreg.

We therefore employ a general specification test due to Hausman (1978) as a formal means of comparison between MLreg and NLSreg, both of which are designed to examine the multivariate relationships between the sensitive trait and respondents' characteristics. The idea is that if the regression modeling assumptions are correct, then NLSreg and MLreg should yield statistically indistinguishable results. If their differences are significant, we reject the null hypothesis of correct specification. Note that model misspecification can arise for various reasons, with measurement error being one possibility. The test statistic and its asymptotic distribution are given by,

$$(\hat{\theta}_{\text{ML}} - \hat{\theta}_{\text{NLS}})^{\top}(\widehat{\mathbb{V}(\hat{\theta}_{\text{NLS}})} - \widehat{\mathbb{V}(\hat{\theta}_{\text{ML}})})^{-1}(\hat{\theta}_{\text{ML}} - \hat{\theta}_{\text{NLS}})^{\top} \overset{\text{approx.}}{\sim} \chi^2_{\dim(\beta)+\dim(\gamma)} \tag{7}$$

where $\hat{\theta}_{\text{NLS}} = (\hat{\beta}_{\text{NLS}}, \hat{\gamma}_{\text{NLS}})$ and $\hat{\theta}_{\text{ML}} = (\hat{\beta}_{\text{ML}}, \hat{\gamma}_{\text{ML}})$ are the NLS and ML estimators and $\widehat{\mathbb{V}(\hat{\theta}_{\text{NLS}})}$ and $\widehat{\mathbb{V}(\hat{\theta}_{\text{ML}})}$ are their estimated asymptotic variances. We view this test as a logical extension of the recommendation in Ahlquist (2017) to informally compare DiM and MLreg.

2.4 Modeling Measurement Error Mechanisms

One advantage of the multivariate regression framework proposed in Imai (2011) is its ability to directly model measurement error mechanisms. For example, Blair and Imai (2012) model floor and ceiling effects. Ahlquist (2017) expresses skepticism with this approach, stating "Trying to address nonstrategic error by making more and stronger assumptions seems like a high-cost, low-return strategy" (p. 5). We disagree. Measurement error models are useful in list experiments just like they are in other settings (see e.g., Carroll et al., 2006). First, these models include the model without measurement error as their limiting case, requiring fewer and weaker assumptions than standard models. As a result, we can apply the specification test as shown for NLSreg and MLreg above. Second, these models can be used to check the robustness of empirical results to measurement error. Third, researchers can use these models to test the mechanisms of survey misreporting in order to understand when list experiments do and do not work.

Although we believe that the top-biased error introduced in Ahlquist (2017) is implausible in practice, we show how to model this error process as an illustration of how our modeling framework can flexibly incorporate various measurement error mechanisms. We then show how to model uniform error, discussed but not examined in detail in Ahlquist (2017), in which "a respondent's truthful response is replaced by a random uniform draw from the possible answers available to her, which in turn depends on her treatment status" (p. 5).[1] We think that this uniform response error

[1] Ahlquist (2017) writes down a linear model in footnote 3 but this generative model is inconsistent with list

process is more realistic and hence the proposed uniform error model will be useful for applied researchers. As shown in Appendix A, DiM is biased under these error processes.

Top-biased error. Under top-biased error, for the NLS estimation, equation (2) becomes,

$$\mathbb{E}(Y_i \mid T_i, X_i) \;=\; pJ + T_i\{p + (1-p)\mathbb{E}(Z_i \mid X_i)\} + (1-p)\mathbb{E}(Y_i^* \mid X_i) \tag{8}$$

where p is the additional parameter representing the population proportion of respondents who give the maximal value as their answer. When $p = 0$ the model reduces to the standard model given in equation (2). The NLS estimator is obtained by minimizing the sum of squared error,

$$(\hat{\beta}_{\text{NLS}}, \hat{\gamma}_{\text{NLS}}) \;=\; \arg\min_{(\beta,\gamma,p)} \sum_{i=1}^{N} [Y_i - pJ - T_i\{p + (1-p)\mathbb{E}(Z_i \mid X_i)\} - (1-p)\mathbb{E}(Y_i^* \mid X_i)]^2 . \tag{9}$$

We can also model top-biased error using the following likelihood function,

$$\begin{aligned}
&\prod_{i \in \mathcal{J}(1,J+1)} \left[g(X_i;\beta) f(X_i;\gamma)^J + p\{1 - g(X_i;\beta) f(X_i;\gamma)^J\}\right] \prod_{i \in \mathcal{J}(0,J)} \left[f(X_i;\gamma)^J + p\{1 - f(X_i;\gamma)^J\}\right] \\
&\prod_{i \in \mathcal{J}(1,0)} (1-p)\{1 - g(X_i;\beta)\}\{1 - f(X_i;\gamma)\}^J \prod_{y=0}^{J-1} \prod_{i \in \mathcal{J}(0,y)} (1-p)\binom{J}{y} f(X_i;\gamma)^y \{1 - f(X_i;\gamma)\}^{J-y} \\
&\prod_{y=1}^{J} \prod_{i \in \mathcal{J}(1,y)} (1-p) \left[g(X_i;\beta)\binom{J}{y-1} f(X_i;\gamma)^{y-1} \{1 - f(X_i;\gamma)\}^{J-y+1} + \right. \\
&\qquad \left. \{1 - g(X_i;\beta)\}\binom{J}{y} f(X_i;\gamma)^y \{1 - f(X_i;\gamma)\}^{J-y} \right]
\end{aligned} \tag{10}$$

Again, when $p = 0$, this likelihood function reduces to the likelihood function of the original model, which is given on the logarithmic scale in equation (6). While this likelihood function is too complex to optimize, we can use the EM algorithm (Dempster, Laird and Rubin, 1977) to maximize it. The details of this algorithm are given in Appendix B.1.

Uniform error. Under the uniform error mechanism, we modify the regression model given in equation (2) to the following,

$$\begin{aligned}
\mathbb{E}(Y_i \mid T_i, X_i) \;=\; & \frac{p_0(1-T_i)J}{2} + T_i \left\{ \frac{p_1(J+1)}{2} + (1-p_1)\mathbb{E}(Z_i \mid X_i) \right\} \\
& + \{(1-T_i)(1-p_0) + T_i(1-p_1)\}\mathbb{E}(Y_i^* \mid X_i)
\end{aligned} \tag{11}$$

where $p_t = \Pr(S_i \mid T_i = t)$ represents the proportion of misreporting individuals under the treatment condition $T_i = t$. Again, when $p_0 = p_1 = 0$, this model reduces to the original model without measurement error. As before, we can obtain the NLS estimator by minimizing the sum of squared error. We can also formulate the ML estimator using the following likelihood function,

$$\prod_{i \in \mathcal{J}(1,J+1)} \left\{ (1-p_1) g(X_i;\beta) f(X_i;\gamma)^J + \frac{p_1}{J+2} \right\} \prod_{i \in \mathcal{J}(1,0)} \left\{ (1-p_1)\{1 - g(X_i;\beta)\}\{1 - f(X_i;\gamma)\}^J + \frac{p_1}{J+2} \right\}$$

experiments as the response can take a non-integer value.

$$\prod_{y=0}^{J} \prod_{i \in \mathcal{J}(0,y)} \left\{ (1-p_0) \binom{J}{y} f(X_i;\gamma)^y \left\{1 - f(X_i;\gamma)\right\}^{J-y} + \frac{p_0}{J+1} \right\}$$

$$\prod_{y=1}^{J} \prod_{i \in \mathcal{J}(1,y)} \left[(1-p_1) \left\{ g(X_i;\beta) \binom{J}{y-1} f(X_i;\gamma)^{y-1} \left\{1 - f(X_i;\gamma)\right\}^{J-y+1} + \right.\right.$$

$$\left.\left. \{1 - g(X_i;\beta)\} \binom{J}{y} f(X_i;\gamma)^y \left\{1 - f(X_i;\gamma)\right\}^{J-y} \right\} + \frac{p_1}{J+2} \right] \tag{12}$$

As shown in Appendix B.2, the EM algorithm can be used to obtain the ML estimator.

2.5 Robust Multivariate Regression Models

Through simulation studies, Ahlquist (2017) criticizes MLreg by pointing out that the estimated proportion of the sensitive trait can be biased in the presence of measurement error. Again, we advise that, if the goal is to estimate the proportion of the sensitive trait, researchers should use DiM rather than multivariate regression models, which are designed to analyze the association between the latent sensitive trait and observed covariates. As discussed in Section 2, doing this requires additional assumptions, for example about the form of the regression function. When only the prevalence of the sensitive trait is of interest, DiM estimator remains the most robust option.

Nevertheless, we show here how to conduct multivariate regression analysis while ensuring that the estimated proportion of sensitive trait is close to DiM. To do this, we follow Chou, Imai and Rosenfeld (2017), who show how to incorporate available auxiliary information such as the aggregate prevalence of sensitive traits when fitting regression models. The authors find that supplying aggregate truths significantly improves the accuracy of list experiment regression models. Following this strategy, we fit the multivariate regression models such that they give the overall prediction of sensitive trait prevalence consistent with DiM. To the extent that DiM rests on weaker assumptions, this modeling strategy may improve the robustness of the multivariate regression models.

Specifically, we use the following additional moment condition,

$$\mathbb{E}\{g(X_i;\beta)\} \quad = \quad \mathbb{E}\left\{ \frac{\sum_{i=1}^N T_i Y_i}{\sum_{i=1}^N T_i} - \frac{\sum_{i=1}^N (1-T_i) Y_i}{\sum_{i=1}^N (1-T_i)} \right\} \tag{13}$$

For the NLS estimation, we combine this moment condition with the following first order conditions from the two-step NLS estimation.

$$\mathbb{E}\left[T_i \{Y_i - f(X_i;\gamma) - g(X_i;\beta)\} g'(X_i;\beta)\right] \quad = \quad 0 \tag{14}$$

$$\mathbb{E}\left[(1-T_i) \{Y_i - f(X_i;\gamma)\} f'(X_i;\gamma)\right] \quad = \quad 0 \tag{15}$$

where $f'(X_i;\gamma)$ and $g'(X_i;\beta)$ are the gradient vector with respect to γ and β, respectively. Altogether, the moments form a generalized method of moments (GMM) estimator. In fact, we can use

the same exact set up Chou, Imai and Rosenfeld (2017) and use their code in the `list` package in R to obtain the NLS estimator with this additional constraint, although the standard errors must be adjusted as the auxiliary constraint does not provide additional information.

We can also incorporate the constraint in equation (13) in the ML framework. To do this, we combine the score conditions obtained from the log-likelihood function with this additional moment condition. Then, the GMM estimator can be constructed using all the moment conditions. The details of this approach are given in Appendix B.3.

3 Revisiting the Simulation Studies in Ahlquist (2017)

As discussed in Section 2.2, we do not believe that top-biased error is plausible. As such, in our view, the simulation studies presented in Ahlquist (2017) are not informative about the performance of multivariate regression models in practice. We provide evidence for this claim in Section 4, where we show that adjusting for top-biased error does not remove the bias found in Ahlquist (2017). Nevertheless, here we apply the various multivariate regression estimators and specification tests described in Section 2 to the simulation settings in Ahlquist (2017) to examine whether the aforementioned theoretical expectations are consistent with simulation results.

Ahlquist (2017) finds that MLreg is more sensitive to top-biased error than DiM. The paper also reports that the degree of sensitivity increases when the prevalence of sensitive trait is low and the control items are negatively correlated with each other. Below, we show that our statistical test, developed in Section 2.3, readily detects the severely misspecified data-generating process used in Ahlquist (2017). We show that NLSreg is robust to these types of model misspecification. Although we confirm that MLreg is sensitive to top-biased error, we find that it is more robust to uniform error. Finally, we find that the new ML estimators proposed above perform reasonably well in the presence of response errors, especially when the sensitive trait is sufficiently common.

3.1 Simulation Settings

We begin by replicating the "designed list" simulation scenario, which Ahlquist (2017) found to be most problematic for MLreg.[2] In addition to the introduction of top-biased error, this simulation scenario violates the assumptions of MLreg in two ways. First, Ahlquist (2017) follows the advice of Glynn (2013) and generate a negative correlation among the control items. By contrast, MLreg assumes conditional independence of the control items. Second, control items are generated with

[2]We do not study his "Blair-Imai list" simulation scenario. Despite what the label suggests, the data generating process used for this simulation does not follow the binomial distribution assumed for MLreg of Blair and Imai (2012), making it impossible to determine the degree to which measurement error causes bias.

different marginal probabilities, which is also inconsistent with the binomial distribution.[3]

The data generating process for these problematic "designed" lists is as follows. For the control outcome Y_i^*, the marginal probabilities are fixed for each control item. For the simulations with $J = 3$ control items, the probabilities of latent binary responses are specified to be $(0.5, 0.5, 0.15)$, whereas in the simulations with $J = 4$, the study uses $(0.5, 0.5, 0.15, 0.85)$. The `rmvbin()` function in the R package `bindata` is used to generate the latent responses to the control items such that the correlation between the first two items is negative 0.6. To generate the sensitive trait, Z_i, first a single covariate, X_i, is sampled independently from the uniform distribution for each observation $i = 1, 2, \ldots, N$. Together with an intercept, we form the model matrix $\mathbf{X}_i = (1, X_i)$. The sensitive trait is then drawn according to the logistic regression given in equation (3). The coefficients are set to $\beta = (0, -4)$ corresponding to the prevalence of sensitive trait approximately equal to 0.12. Finally, we assign the half of the sample to the treatment group ($T_i = 1$) and the other half to the control group ($T_i = 0$). The outcome variable then is generated according to equation (1).

To introduce top-biased error, Ahlquist (2017) uses complete randomization to select 3% of the sample and changes the outcome variable Y_i to $J + 1$ (J) if observation is assigned to the treatment (control) group, independently of the values of $\mathbf{X}_i$, Y_i^*, and Z_i. To generate uniform error, we similarly sample 3% of the observations and assign their outcome variable with uniform probability to one of the $J + 2$ ($J + 1$) possible values, depending on their treatment status. We follow these procedures in our simulations.

3.2 Detecting the Model Misspecification

Given this radically different data generating process, it is unsurprising that Ahlquist (2017) finds MLreg to be severely biased. However, as explained in Section 2.1, NLSreg should be more robust than MLreg though it is less efficient. This bias-variance tradeoff arises because NLSreg does not assume the binomial distribution for the control items. Under the assumptions of NLSreg, the control items can be arbitrarily correlated and have different marginal probabilities (although a specific functional form – here, the logistic function – is assumed for the conditional expectation). This implies that NLSreg should only be subject to the potential bias from response error.

This theoretical expectation suggests that the Hausman test proposed in Section 2.3 may be able to detect the departure from the modeling assumptions. We find that this is indeed the case. Table 2 shows that our test diagnoses the inconsistency of MLreg in the presence of such severe model misspecification. The table presents the rejection rate for our simulations at different combinations

[3] Blair and Imai (2012) show how to model this data generating process using the poisson binomial distribution. Another possibility is to model the joint distribution of control items by using the information from another survey, in which each item is asked separately.

Number of control items	Sample size	No response error p-value	No response error p-value & negative	Top-biased error p-value	Top-biased error p-value & negative	Uniform error p-value	Uniform error p-value & negative
J = 3	1000	0.10	0.71	0.44	0.83	0.65	0.92
	1500	0.07	0.67	0.47	0.80	0.80	0.96
	2000	0.03	0.62	0.50	0.80	0.90	0.98
J = r	1000	0.17	0.84	0.52	0.88	0.62	0.92
	1500	0.21	0.91	0.59	0.88	0.82	0.95
	2000	0.19	0.95	0.62	0.88	0.90	0.97

Table 2: Results of the Model Misspecification Test for the Designed List Simulations. The proportions rejecting the null hypothesis of no model misspecification are shown. The "p-value" column is based on the standard Hausman test with p-value of 0.05 as the threshold, while the "p-value & negative" column is based on the combined rejection criteria where we reject the null hypothesis if p-value is less than 0.05 or the test statistic takes a negative value.

of J and N with p-value of 0.05 as the threshold. As the "p-value" columns show, we find sufficiently large (and positive) test statistics to reject the null hypothesis of no misspecification in a large proportion of trials, especially when there is no response error. The finding is consistent with the gross model misspecification, in excess of response error, introduced by the designed list procedure in Ahlquist (2017). Interestingly, the top-biased error appears to mask this misspecification.

Importantly, we discovered that in all cases a large proportion of the trials yielded a *negative* value of the test statistic, which correspond to extremely poor fit of MLreg relative to NLSreg. Such values are only consistent with model misspecification so long as the test is sufficiently powered (Schreiber, 2008). While the test statistic can by chance take a negative value in a finite sample, in our simulations such statistics are strikingly prevalent. As shown in the "p-value & negative" columns of the table, by using a negative or large positive test statistic as the criterion for rejection, we obtain a much more powerful test for misspecification even in the case of top-biased error. Although this test may be conservative, leading to over-rejection of the null hypothesis, the fact that these rejection rates are large even when the sample size is moderate suggests that the test is well-powered in this simulation study.

In sum, the proposed model misspecification test readily diagnoses the designed list misspecification studied in Ahlquist (2017). This finding is a corrective to Ahlquist (2017), which includes the model specification test as one of the "strategies that will *not* work" (italics in original; p. 15). Although the test may not work in all settings, we find here that it detects the significant problems in the designed lists simulations at a high rate.

3.3 Robustness of the Nonlinear Least Squares Regression Estimator

Our second main finding is that NLSreg is robust to these misspecifications. This fortifies our previous result that the model misspecification test, being based on the divergence between NL-

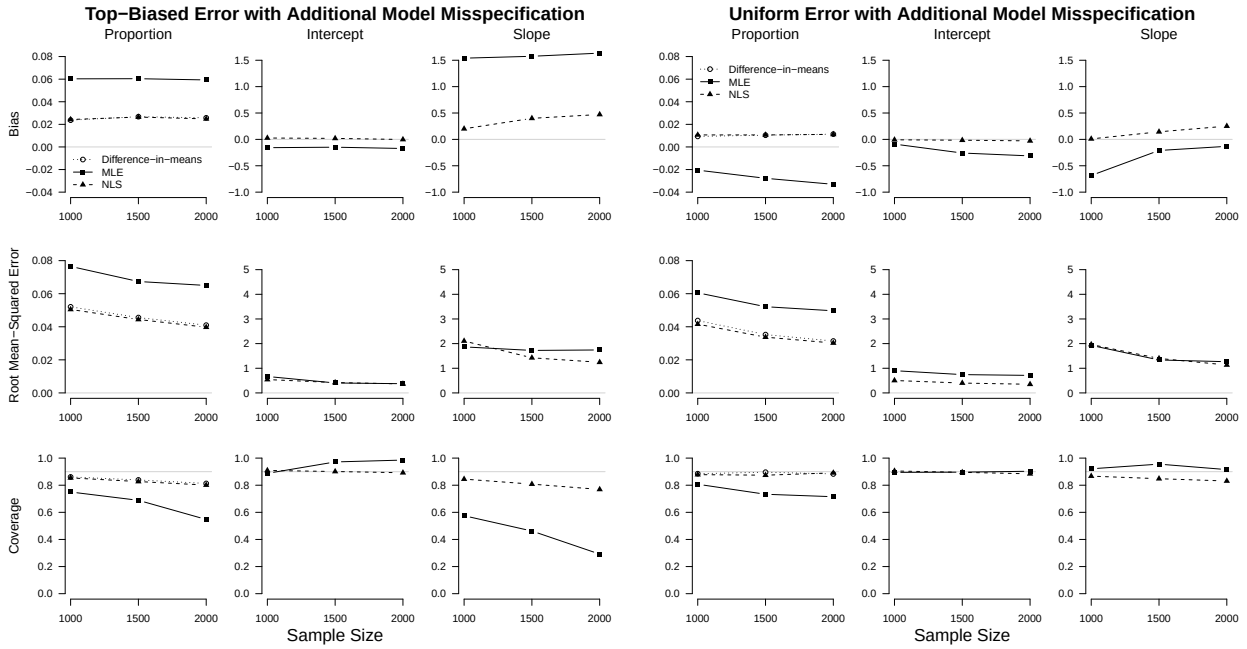

Figure 1: Robustness of the Nonlinear Least Squares Regression Estimator in the Presence of Several Model Misspecifications Considered in Ahlquist (2017). We consider the three estimators of the prevalence of sensitive trait: the difference-in-means estimator (open circle with dotted line), the maximum likelihood (ML) regression estimator (solid square with solid line; filtered by the model misspecification test using the combined criteria), and the nonlinear least squares (NLS) estimator (solid triangle with dashed line). The result shows that the NLS regression estimator is as robust as the difference-in-means estimator.

Sreg and MLreg, is effective for diagnosing model misspecification. To illustrate the robustness of NLSreg, in Figure 1, we present the bias, root-mean-squared-error (RMSE), and the coverage of 90% confidence intervals of our estimates of the population prevalence of the sensitive trait. We also present results for MLreg, filtered using the p-value plus negative value criterion for rejection described above.[4] Lastly, given the goals of multivariate regression, we also compute these statistics for the estimated coefficients.

Figure 1 shows the results for $J = 3$ control items for top-biased error (left three columns) and uniform error (right three columns). For the space limitation, the analogous figure for $J = 4$ control items is shown in Figure 6 in Appendix C. We include the three estimators considered in Ahlquist (2017): DiM, MLreg (solid square with solid line), and NLSreg (solid triangle with dashed line). Our main finding here is that NLSreg is robust to all of these model misspecifications, doing as well as DiM. This is consistent with our prior expectation: DiM is a special case of NLSreg.

[4]For the purpose of presentation, we adopt the approach suggested by Blair and Imai (2012) to introduce weakly informative priors to the sensitive item logistic regression to address complete separation of covariates in a small number of models affected by model misspecification (Gelman et al., 2008). Although fewer than 1% of simulations are affected by separation issues, they drive up the bias and RMSE, making it difficult to present them graphically if one does not use regularization. The findings remain qualitatively unaffected by this choice.

Although filtering based on the model misspecification test addresses the overestimation of the sensitive trait under top-biased error observed in Ahlquist (2017), we note that MLreg does not perform well for the estimation of the coefficients, nor does it improve inference for the prevalence of the sensitive trait under uniform error. However, as Table 2 showed, these results are based on the small fraction of trials that did not result in a negative or large positive test statistic. In such trials, the NLS estimates were also inaccurate due to sampling error. This suggests that, while our proposed statistical test will often be able to detect misspecification in practice, in the instances where it does not, NLSreg (and, by extension, DiM) will also be biased.

The results confirm our theoretical expectation that NLSreg is robust to various types of misspecification. As a final point, we note that our simulation results, based on the grossly misspecified data generating process used in Ahlquist (2017), do not imply that MLreg will always perform badly for designed lists. Although all models make simplifying assumptions, we find the simulation procedure used in Ahlquist (2017) to be much too artificial. It implies, for example, that the individual covariates will not be predictive of the control items whatsoever. This is rarely if ever the case with real-world data. Furthermore, our model misspecification test is able to detect the distributional misspecification in these simulations, suggesting that, in practice, such a misspecification will often be apparent from the divergence of the NLS and ML results.

3.4 Addressing Response Error

As shown above, the simulation settings used in Ahlquist (2017) and replicated above include several violations of the modeling assumptions, including correlation between control items, varying control item propensities, model misspecification, and measurement error. As such, it is difficult to disentangle which model misspecifications, or combination of them, are affecting the performance of different methods. In this section, we focus on assessing the impacts of top-biased and uniform error processes and examine how the multivariate models proposed in Section 2 can address them. To be sure, applied researchers will rarely know which (if any) of these mechanisms characterize their data. Nevertheless, we show here that these methods can eliminate these errors when they arise, and illustrate in Section 4 how they can be used to assess the robustness of empirical results.

To isolate the effects of measurement errors, we develop a data generating process that assumes no other model misspecification. First, we draw the latent response to the sensitive item Z_i and the control item Y_i^* according to the model defined in equations (3) and (4), we introduce response error. Following Ahlquist (2017), we set the true values of the coefficients for the control items to γ to $(0, 1)$, corresponding to a conditional mean of observed response about $J \times 0.62$, whereas the coefficients for the sensitive item are set to $\beta = (0, -4)$, generating a low propensity of approximately 0.12. In Appendix C, we present a high propensity scenario of about 0.38. Lastly,

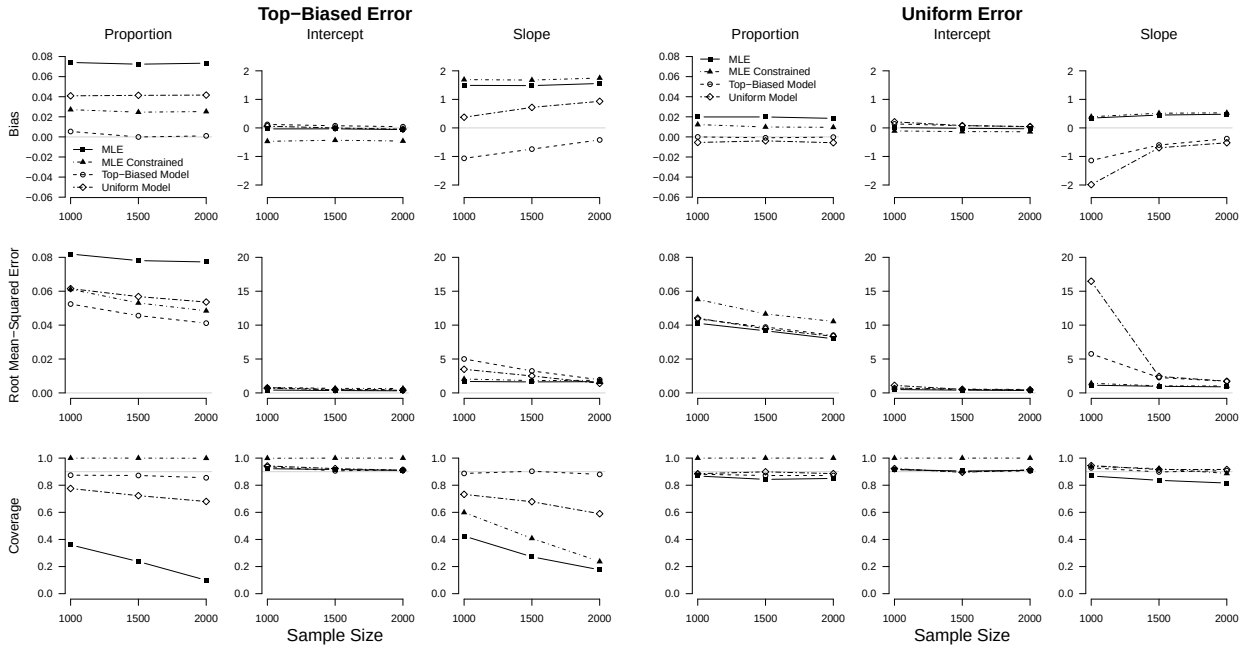

Figure 2: Robustness of the Constrained and Measurement Error Maximum Likelihood Estimators in the Presence of Response Errors when the Propensity of Sensitive Trait is Low. We consider four estimators of the prevalence of sensitive trait and slope and intercept regression coefficients: the standard maximum likelihood (ML) estimator (solid square with solid line), the constrained ML estimator (solid triangle with dot-dash line), the ML estimators adjusting for top-biased response error (open circle with dashed lines) and uniform response errors (open diamond with dot-long-dash line). The result shows that both the constrained MLE estimator and the models adjusting for response error are an improvement over the performance of the MLE estimator.

we then introduce each response error using the corresponding procedure described earlier.

Figure 2 presents the findings for $J = 3$ whereas Figure 7 of Appendix C shows the results for $J = 4$. In the left-hand columns, we show the top-biased error simulation with the standard ML estimator (solid square with solid line), the constrained ML estimator (solid triangle with dot-dash line), the top-biased ML model (open circle with dash line), and the uniform ML model (open diamond with dot-long-dash line) introduced in Section 2. The right-hand columns presents the uniform error simulation results. As before, we show the bias, RMSE, and coverage of 90% confidence intervals.

As the upper left-hand corner plot shows, we are able to replicate the main finding in Ahlquist (2017) that a small amount of top-biased error is enough to significantly bias the standard ML estimator. Looking at the regression coefficients, we find that this positive bias follows from the bias in the estimated slope. Our proposed methods appear to address this issue effectively. The constrained estimator slashes the bias of the overall prevalence by almost 75%. This is unsurprising, as it constrains the regression-based prediction to the difference-in-means estimate. However, because the constrained ML does not model the error mechanism directly, it does not improve the bias of the estimated regression coefficients. Indeed, the dashed lines show, the constrained model

reduces the bias by decreasing the intercept rather than the slope, which does not help in this particular simulation setting where the bias for the intercept is small to begin with. As a result, the coverage of confidence interval for the slope is only slightly improved.

As expected, the top-biased error model most effectively addresses this measurement error mechanism, eliminating the bias of the three quantities of interest almost entirely. Likewise, coverage is at the nominal rates for all three quantities of interest. We find that the uniform error model, which models a different error process to the one assumed, nevertheless is no worse than the standard ML model. Indeed, it exhibits less bias, better coverage, and lower RMSE than MLreg. In both cases, there is a small finite-sample bias that reduces as sample size increases.

The right-hand columns of Figure 2 examine the performance of the same four estimators under uniform error. We find several interesting results. Most importantly, we find that MLreg is significantly less biased under this measurement error mechanism than under the top-biased error process. Given the greater plausibility of uniform error relative to top-biased error, this finding suggests that MLreg may be more robust to nonstrategic measurement error than the simulations of Ahlquist (2017) suggest. Indeed, in our empirical reanalysis (see Section 4), we find that MLreg performs well for the list experiment on texting while driving, despite the observation in Ahlquist (2017) that many respondents appeared to rush through the survey haphazardly

We find that the uniform error model leads to some under-estimation of the sensitive trait prevalence. While no estimator is biased for estimating the intercept, the uniform error model yields a large finite-sample bias for the estimation of slope coefficient. However, these biases are small relative to the standard error as shown by the proper coverage of the corresponding confidence intervals, and they go to zero as the sample size increases. In contrast, the constrained ML estimator appears to perform well with small bias and RMSE as well as proper coverage of confidence intervals. We also note that the top-biased error model, which assumes a different error process than the simulation DGP, performs well under uniform error exhibiting low bias, RMSE, and nominal coverage.

4 Revisiting the Empirical Applications in Ahlquist (2017)

In this section, we revisit a set of list experiments marshaled in Ahlquist (2017) to critique MLreg for list experiments. These experiments were originally reported in Ahlquist, Mayer and Jackman (2014) for the purpose of measuring voter impersonation in the 2012 US election – a phenomenon that many scholars of American politics consider to be exceedingly rare (see, e.g., Sobel, 2009, and references therein). While the difference-in-means estimate from the voter fraud experiment, negative 1.2%, confirms this prior expectation, Ahlquist (2017) finds that the multivariate regression model significantly overestimates voter fraud. Ahlquist, Mayer and Jackman (2014) also

conducted two additional list experiments, one on alien abduction and the other on texting while driving. Ahlquist (2017) finds that MLreg similarly overestimates the prevalence of alien abduction, while no such problem is found for the texting-while-driving list.

Below, we reanalyze these list experiments using the proposed methodology. As a preliminary point, we show that a simple descriptive analysis of these list experiments can indicate that the application of multivariate regression models to the voter fraud and alien abduction lists is inappropriate. Our analysis confirms that these are extremely rare or non-existent events, and consequently no association exists to be studied. Nevertheless, our reanalysis of these data produces more sensible estimates of the prevalence of voter fraud and alien abduction. In particular, when accounting for uniform error the estimated prevalence of these events is precisely zero. Finally, we analyze the texting-while-driving list, which measures a much more common event, and show that the proposed methods as well as the standard ML estimator yield reasonable results.

4.1 Extremely Rare Sensitive Traits and Multivariate Regression Analysis

We begin by cautioning against the use of multivariate regression models when studying extremely rare or even nonexistent sensitive traits. The reason is simple. The goal of multivariate regression analysis is to measure the association between sensitive traits and respondent characteristics. If almost all respondents do not possess such traits, as is the case for voter impersonation in the US and alien abduction, then multivariate regression analysis is likely to be unreliable because no association exists in the first place (and any existing association is likely to be due to noise). Therefore, it is no surprise that Ahlquist (2017) finds the ML regression estimator to be misleading for the list experiments on voter fraud and alien abduction but unproblematic for the list experiment on texting while driving, which is known to be much more common than the other two phenomena.

Indeed, we find that the voter fraud and alien abduction list experiments elicit extremely small proportions of the affirmative answer. As discussed in Section 2.3 and recommended in Blair and Imai (2012), Table 3 presents the estimated proportion of each respondent $\mathcal{J}(y, z)$ type defined by two latent variables, i.e., the total number of affirmative answers to the control items $Y_i^* = y$ and the answer to the sensitive item $Z_i = z$. We also present DiM for each list experiment. The list experiment on voter fraud is most problematic, yielding an overall negative estimate and exhibiting three negative estimates for respondent types who would answer the sensitive item affirmatively.

Although these negative estimates are not statistically significant, this simple table implies that the list experiment on voter fraud suffers from either the violation of assumptions or an exceedingly small number of respondents with the sensitive trait. The descriptive information clearly suggests that multivariate regression analysis is not appropriate for the list experiments on voter fraud and alien abduction. There is almost no respondent who would answer yes to these questions, and as a

	Voter fraud		Alien abduction		Texting while driving	
	est.	s.e.	est.	s.e.	est.	s.e.
$\mathcal{J}(0,\,1)$	−0.015	0.015	0.004	0.017	0.034	0.015
$\mathcal{J}(1,\,1)$	−0.020	0.017	0.007	0.014	0.087	0.016
$\mathcal{J}(2,\,1)$	−0.008	0.012	0.016	0.009	0.047	0.012
$\mathcal{J}(3,\,1)$	0.004	0.009	0.011	0.006	0.022	0.008
$\mathcal{J}(4,\,1)$	0.027	0.004	0.024	0.004	0.033	0.005
$\mathcal{J}(0,\,0)$	0.232	0.011	0.348	0.012	0.217	0.011
$\mathcal{J}(1,\,0)$	0.469	0.016	0.467	0.016	0.419	0.016
$\mathcal{J}(2,\,0)$	0.204	0.015	0.106	0.012	0.104	0.014
$\mathcal{J}(3,\,0)$	0.070	0.011	0.012	0.008	0.032	0.010
$\mathcal{J}(4,\,0)$	0.037	0.008	0.004	0.006	0.005	0.007
Diff.-in-means	-0.012	0.041	0.062	0.036	0.223	0.039

Table 3: Estimated Proportion of Respondent Types by the Number of Affirmative Answers to the Control and Sensitive Items. The table shows the estimated proportion of respondent type $\mathcal{J}(y, z)$, where $y \in \{0, \ldots, J\}$ denotes the number of affirmative answers to the control items and $z \in \{0, 1\}$ denotes whether respondents would answer yes to the sensitive item. In the last row, we also present the difference-in-means estimator for the estimated proportion of those who would affirmatively answer the sensitive item. The voter fraud and alien abduction list experiments have an extremely small proportion of those who would answer yes to the sensitive item, and for voter fraud some proportions are estimated negative, suggesting the problem of list experiment.

result, there is no association to be studied. Therefore, the only list experiment that would sensibly benefit from multivariate regression analysis is the one on texting while driving.

The application of the multivariate ML regression model in Ahlquist (2017) to the alien abduction and voter fraud lists compounds the weakness of indirect questioning methods, which are poorly suited to studying extremely rare sensitive traits. Although indirect questioning methods seek to reduce bias from social desirability and nonresponse by partially protecting the privacy of respondents, they are much less efficient than direct questioning. As a consequence, the estimates will lack the statistical precision required for measurement and analysis of extremely rare traits. Indirect methods also amplify finite sample bias associated with rare events (King and Zeng, 2001).

Given these tradeoffs, we recommend that list experiments be used only when studying truly sensitive topics. The increased cognitive burden on respondents and the loss of statistical efficiency are too great for this survey methodology to be helpful for non-sensitive traits. Among the three list experiments, the one on alien abduction provides the smallest insight into the efficacy of list experiments. In fact, such questions may increase measurement error if respondents stop taking the survey seriously. Better designed validation studies are needed to evaluate the effectiveness of list experiments and their statistical methods (see e.g., Rosenfeld, Imai and Shapiro, 2016).

Despite our reservations about the application of the multivariate regression models to two of the three list experiments, we now examine whether the methods proposed in Section 2 can

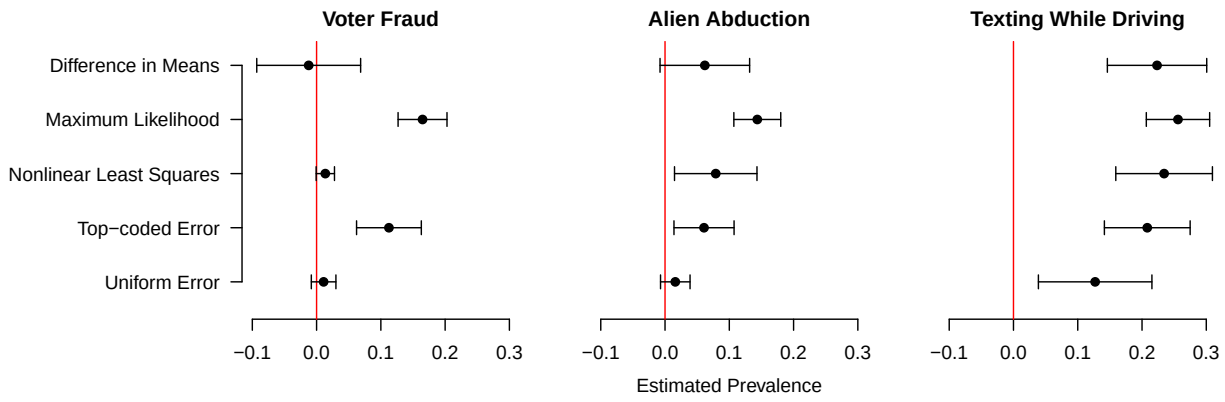

Figure 3: The Estimated Prevalence of Voter Fraud, Alien Abduction, and Texting while Driving. Along with the difference-in-means estimator, we present the estimates based on various multivariate regression analyses including the maximum likelihood and nonlinear least squares regression estimators and estimates are much more stable. The results based on the two measurement error models, i.e., top-biased and uniform errors, are also presented.

detect and adjust for measurement error by applying them to these list experiments. We repeat, however, that although the focus in Ahlquist (2017) is on estimating the proportion of respondents who possess a sensitive trait, these regression models have been developed to enable the statistical analysis of the association between sensitive traits and respondents' characteristics. If the goal of analysis is to estimate the prevalence of sensitive traits, then the difference-in-means estimator is a simple and robust solution and there is no need to use multivariate regression models.

4.2 Comparison between the ML and NLS Regression Estimators

We conduct two types of analyses. First, we fit the multivariate regression model with age, gender, and race as covariates using the ML and NLS estimation methods. Below, we find that the NLS regression estimator, which was not examined in Ahlquist (2017), is robust to the problems identified in the original article. We also show that this diference between MLreg and NLSreg can be exploited to detect model misspecification as explained in Section 2.3. Our approach provides a principled alternative to the suggestion in Ahlquist (2017) to simply compare DiM and MLreg.

We begin by showing that NLSreg is more robust to the overestimation issue identified in Ahlquist (2017). Figure 3 presents the estimated proportion of sensitive trait based on DiM, NLSreg, and MLreg as well as two other estimators that will be described later. We find that the NLS regression estimates closely track the difference-in-means estimates. Indeed, the NLS estimate is statistically indistinguishable from the difference-in-means estimate in all three cases. In the case of voter fraud, the NLS estimate exceeds the difference-in-means estimate by 2.59 percentage points, with a 90% bootstrap confidence interval equal to $[-4.06, 9.96]$. The resulting NLS estimate of voter fraud is around 1.4% and statistically indistinguishable from 0. For the alien abduction list

	degrees of freedom	test statistic	p-value
Voter fraud			
Without covariates	2	−0.001	0.99
With covariates	8	−29.215	<0.01
Alien abduction			
Without covariates	2	−13.283	<0.01
With covariates	8	14.961	0.06
Texting while driving			
Without covariates	2	0.662	0.72
With covariates	8	2.367	0.97

Table 4: Results of the Proposed Specification Tests. The p-values are based on the absolute value of the test statistic. The results show that for the list experiments based on voter fraud and alien abduction we detect model misspecification. For the list experiment on texting while driving, we fail to reject the null hypothesis.

experiment, the NLS estimate exceeds the difference-in-means estimate by 1.70 percentage points, also an insignificant difference (90% CI: $[-0.59, 6.78]$). Lastly, for the texting-while driving list experiment, the difference is 1.10 percentage points (90% CI: $[-1.12, 2.63]$.)

Having shown that NLSreg does not yield meaningfully larger estimates than DiM, we now apply the statistical test developed in Section 2.3 to these list experiments. Table 4 presents the results of this proposed specification test. For the list experiments on voter fraud and alien abduction, which our earlier analysis found most problematic, we obtain negative and large, positive values of the Hausman test statistic. For the negative test statistic, we compute p-values based on the absolute value. The results of the statistical hypothesis tests strongly suggest that the model is misspecified for the voter fraud and alien abduction experiments. In contrast, for the list experiment on texting while driving, we fail to reject the null hypothesis of correct model specification.

In sum, the proposed model specification test reaches the same conclusion as the descriptive analysis presented above: multivariate regression models are not appropriate for list experiments with extremely rare sensitive traits. For the list experiments on voter fraud and alien abduction, we find strong evidence for model misspecification, suggesting that the results of multivariate regression analysis will be unreliable for these list experiments. In contrast, we fail to find such evidence for the list experiment on texting while driving, for which the proportion of affirmative answers appears to be relatively high.

4.3 Modeling Response Error

Next, we apply the nonstrategic measurement error models developed in Section 2.4 to these data sets and examine whether they yield different results. Earlier, we have argued that the top-biased error process makes little sense because it implies that respondents are willing to reveal sensitive traits. We would expect top-biased error to be most unlikely for the list experiment on voter fraud,

as this is the most unambiguously stigmatized trait. On the other hand, uniform error may be the more plausible measurement error mechanism due for example to satisficing.

As shown in Figure 3, the results based on these measurement error models are consistent with our arguments. For the list experiment on voter fraud, the top-biased error gives a relatively large estimate that is statistically indistinguishable from the ML estimate. In contrast, the uniform error model provides an estimate that is indistinguishable from zero with the 90% confidence interval that is narrower than any other estimator. The list experiment on alien abduction yields a similar result. Like all other models, the top-biased error model gives an estimate that is statistically distinguishable from zero, suggesting that 6 percent of respondents were abducted by aliens (even the difference-in-means estimate is barely statistically insignificant). On the other hand, the prevalence estimate based on the uniform error process model has the narrowest 90% confidence interval that contains zero, suggesting a superior model fit. The results indicate that the uniform error model is more effective for mitigating nonstrategic respondent error than the top-biased error model.

Finally, the results for the list experiment on texting while driving show that the top-biased and uniform measurement error models yield estimates that are much more consistent with the other models. Although the estimate based on the uniform error model is smaller, it has a wider confidence interval than other estimates, suggesting a possibly poor model fit. Together with the other results shown in this section, this finding implies that only the estimates based on the list experiment on texting while driving are robust to various model specification and measurement errors, whereas the estimates for voter fraud and alien abduction are quite sensitive. In the following analysis, we show that accounting for measurement error does not alter any of the multivariate inferences that one would draw from this list experiment.

A careful statistical analysis like the one shown here reveals that the problems of multivariate regression models described in Ahlquist (2017) are likely to arise from poor implementation of survey experiments as well as the fact that voter fraud in the US and alien abduction are extremely rare or nonexistent events. Given all three lists were implemented by the same researchers on the same sample, and were consequently subject to the same types of nonstrategic measurement error, the robustness of the texting-while-driving lists suggests that researchers should primarily be concerned with the rarity of the traits under study rather than nonstrategic measurement error.

4.4 Multivariate Regression Analysis of Texting While Driving

Recall that the goal of multivariate regression models for list experiments is to measure the association between the sensitive trait and respondent characteristics. We reiterate that voter fraud and alien abduction are so rare that multivariate analysis of these traits is likely to be unreliable. By contrast, the texting-while-driving list offers a unique opportunity to examine how account-

	NLS		ML		Robust ML		Top-biased		Uniform	
	est.	se.	est.	se.	est.	se.	est.	se.	est.	se.
Sensitive Trait										
(Intercept)	0.031	0.550	−0.272	0.305	−0.466	0.437	−0.351	0.414	−0.109	0.620
Age	−0.032	0.017	−0.017	0.008	−0.015	0.009	−0.015	0.013	−0.063	0.034
White	−0.331	0.482	−0.333	0.299	−0.412	0.330	−0.303	0.466	0.290	0.877
Female	−0.325	0.447	−0.186	0.267	−0.175	0.258	−0.765	0.440	−1.508	1.043
Control Items										
(Intercept)	−0.575	0.078	−0.540	0.060	−0.527	0.064	−0.658	0.069	−0.658	0.080
Age	−0.008	0.002	−0.009	0.002	−0.010	0.002	−0.011	0.002	−0.012	0.002
White	−0.204	0.067	−0.217	0.055	−0.208	0.056	−0.169	0.067	−0.216	0.073
Female	0.003	0.062	−0.014	0.049	−0.020	0.049	0.030	0.060	0.077	0.071

Table 5: Multivariate Regression Analysis of the Texting While Driving List. This table shows the estimated coefficients from the baseline non-linear least squares (NLS) and maximum likelihood (ML) multivariate regression models, as well as the proposed robust ML, top-biased, and uniform measurement error models. Younger respondents are more likely to text while driving. We also find suggestive evidence that male respondents are more likely to text while driving.

ing for measurement error affects the estimated regression parameters. Studies commonly assume that younger drivers are especially likely to text while driving. However, frequently used methods, such as analysis of traffic accidents, are unable to measure this association directly (e.g., Delgado, Wanner and McDonald, 2016). Clearly, DiM also fails to shed light on this relationship.

Figure 4 and Table 5 present the results from our multivariate analysis of the texting-while-driving list. In Figure 4, we present predicted values for the different subgroups defined by the three covariates. These values are calculated by changing the variable of interest while fixing the other variables to their observed values. The highlighted comparisons correspond to statistically significant coefficients at the $\alpha = 0.10$ level (see Table 5). As the bottom-left panel of the figure shows, we find a consistent association between age and texting while driving. In all models younger respondents are more likely to text while driving than older respondents. We find some evidence of gender differentiation as well; male respondents appear to be consistently more likely to text while driving than female respondents, although this difference is not generally statistically significant.

The overall conclusion is that accounting for uniform error does not have a substantial effect on the conclusions that one would draw from the standard ML or NLS models. Moreover, the results illustrate the primary advantage of multivariate regression modeling, which is to assess the relationship between the sensitive trait and covariates. If the goal is only to estimate the prevalence of the sensitive trait, one should use DiM, which is simple and robust.

5 Concluding Recommendations

In this paper, we develop statistical tools that researchers can use to detect and mitigate non-strategic measurement error in list experiments, arising for instance from enumerator or respondent

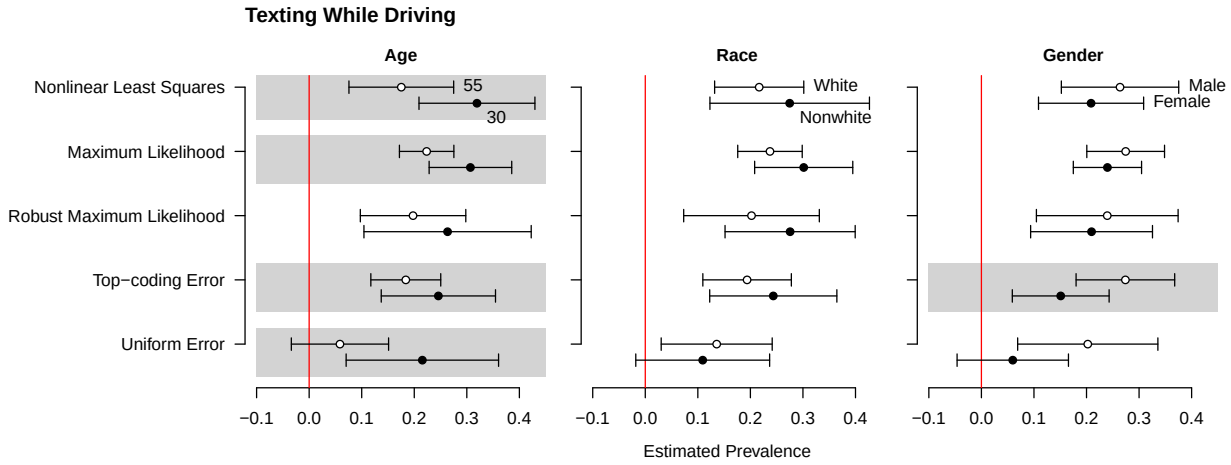

Figure 4: Multivariate Regression Analysis of the Texting While Driving List. This figure shows the estimated prevalence of texting while driving based on the different values of the predictor variables. Highlighted estimates correspond to significant coefficients at the $\alpha = 0.10$ level. Accounting for measurement error, we find that younger respondents are significantly more likely to report texting while driving. We also find that male respondents are more likely to report texting while driving, though the differences are generally insignificant.

noncompliance. At the same time, we agree with Ahlquist (2017) that the best cure for nonstrategic measurement error is to minimize it at the design and administration stages of surveys, and consequently that the presence of such error can signal more fundamental flaws in the survey implementation. For example, because the top-biased error process runs directly against respondents' incentives to conceal sensitive traits, its presence suggests that they do not actually consider the topic to be sensitive in the first place. We would advise against the use of indirect questioning for such topics. The increased burden and variance of indirect questioning are too great to be used for nonsensitive traits. We caution against the use of list experiments for topics like alien abduction, as this can prevent serious engagement with the survey and increase measurement error as a result.

List experiments are motivated by the difficulty of studying truly sensitive topics, which present respondents with strong incentives to conceal their truthful responses. Given such pressures, we believe that the prevailing emphasis on strategic measurement error, such as ceiling and floor effects, is appropriate. For questions that do not pose these strategic dilemmas, list experiments are a burdensome and inefficient alternative to direct questioning.

Despite these and other differences, we agree with Ahlquist (2017) that the consideration of measurement error in the design and analysis of list experiments is important. We conclude this paper by providing seven recommendations about how to analyze list experiments with measurement error.

1. If the goal is to estimate the prevalence of the sensitive trait, researchers should use the difference-in-means estimator. Multivariate regression models should be used only when

inferring the association between the sensitive trait and respondent characteristics.

2. Multivariate regression models should not be used if the the difference-in-means estimator yields a small or negative estimate of the prevalence. A sensitive trait must exist for it to vary with respondent characteristics. In general, list experiments are not suitable for studying rare sensitive traits because they lack statistical power.[5]

3. It is important to first conduct a descriptive analysis as shown in Table 3. In particular, negative estimates of respondent type proportions would suggest that at least one of the identification assumptions of list experiments may have been violated (related statistical tests are described in Blair and Imai (2012) and Aronow et al. (2015)).

4. Researchers should use both the NLS and ML regression estimators. NLSreg relies on weaker assumptions than MLreg, and as a result the former is more robust (though less efficient) than the latter. Despite the greater fragility of MLreg, its reduced variance can matter greatly when analyzing list experiments, already a statistically underpowered questioning mode. To help researchers adjudicate this bias-variance tradeoff, we provide a statistical model misspecification test predicated on the difference between MLreg and NLSreg.

5. Multivariate regression models can be extended to model strategic and nonstrategic measurement error processes. These models can be used as robustness checks. Although many practical steps can be taken to address nonstrategic error, even perfectly administered surveys are subject to strategic measurement error such as ceiling and floor effects. Among nonstrategic measurement error mechanisms, uniform error is more plausible than top-biased error for sensitive topics.

6. It is possible to make the NLS and ML estimators robust by using auxiliary information whenever available. In particular, aggregate truths will be helpful. Even when such information is not available, one can ensure that the NLS and ML regression estimators give the results consistent with the difference-in-means estimator.

7. We agree with Ahlquist (2017) that it is important to use direct question as well as other indirect questioning techniques. The methods developed by Blair, Imai and Lyall (2014) and Aronow et al. (2015) can be used to conduct more credible analyses by combining the data from different survey methods. We also agree with Ahlquist (2017)'s thoughtful recommendations on survey implementation to avoid measurement error by design.

[5] As a corollary, we disagree with the recommendation of Ahlquist (2017) that researchers ask calibration questions involving low-prevalence traits – especially when these may encourage rather than simply gauge measurement error.

References

Ahlquist, John S. 2017. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis*.

Ahlquist, John S., Kenneth R. Mayer and Simon Jackman. 2014. "Alien Abduction and Voter Impersonation in the 2012 U.S. General Election: Evidence from a Survey List Experiment." *Election Law Journal* 13.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford and Donald P. Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3:43–66.

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

Blair, Graeme, Kosuke Imai and Jason Lyall. 2014. "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan." *American Journal of Political Science* 58:1043–1063.

Blair, Graeme, Kosuke Imai and Yang-Yang Zhou. 2015. "Design and Analysis of Randomized Response Technique." *Journal of the American Statistical Association* 110:1304–1319.

Blair, Graeme, Winston Chou and Kosuke Imai. 2017. "list: Statistical Methods for the Item Count Technique and List Experiment." available at the Comprehensive R Archive Network (CRAN). `https://CRAN.R-project.org/package=list`.

Bullock, Will, Kosuke Imai and Jacob N. Shapiro. 2011. "Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan." *Political Analysis* 19:363–384.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. London: Chapman & Hall.

Chou, Winston, Kosuke Imai and Bryn Rosenfeld. 2017. "Sensitive Survey Questions with Auxiliary Information." *Sociological Methods & Research*. Forthcoming.

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

Delgado, M Kit, Kathryn J Wanner and Catherine McDonald. 2016. "Adolescent cellphone use while driving: an overview of the literature and promising future directions for prevention." *Media and communication* 4:79.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B, Methodological* 39:1–37.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics* 2:1360–1383.

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

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

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46:1251–1271.

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

King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9:137–163.

Lyall, Jason, Graeme Blair and Kosuke Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107:679–705.

Rosenfeld, Bryn, Kosuke Imai and Jacob Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60:783–802.

Schreiber, Sven. 2008. "The Hausman Test Statistic Can Be Negative, Even Asymptotically." *Jahrbücher für Nationalökonomie und Statistik* 228:394–405.

Sobel, Richard. 2009. "Voter-ID Issues in Politics and Political Science: Editor's Introduction." *PS: Political Science & Politics* 42:81–85.

A The Bias of the Difference-in-Means Estimator under Non-strategic Measurement Error

In this appendix, we show that the difference-in-means estimator is generally biased under the top-biased and uniform error processes. In both cases, the difference-in-means estimator is generally biased because the range of the response variable (and therefore the magnitude of the measurement error bias) is correlated with the treatment status. In particular, bias is large when the prevalence of sensitive trait is small.

First, under the top-biased error process, the bias of the difference-in-means estimator is given by:

$$\{\mathbb{E}\left[(1-p)(Y_i^* + Z_i) + p(J+1)\right] - \mathbb{E}\left[(1-p)Y_i^* + pJ\right]\} - \tau \;\; = \;\; (1-p)\tau + p - \tau \;\; = \;\; p(1-\tau)$$

where $\tau = \Pr(Z_i = 1)$ is the proportion of those with a sensitive trait, p is the proportion of those who answer $J+1$ regardless of truthful response. The result shows that the bias is zero only when $\tau = 1$, i.e., everyone has a sensitive trait. The bias increases as the prevalence of sensitive trait decreases. Similarly, under the uniform measurement error mechanism, the bias is given by,

$$\left\{\mathbb{E}\left[(1-p)(Y_i^* + Z_i) + p\frac{J+1}{2}\right] - \mathbb{E}\left[(1-p)Y_i^* + p\frac{J}{2}\right]\right\} - \tau \;\; = \;\; (1-p)\tau + \frac{p}{2} - \tau \;\; = \;\; p\left(\frac{1}{2} - \tau\right)$$

Thus, in this case, the bias disappears only when the proportion of those with a sensitive trait exactly equals 0.5. Again, the bias is large when the prevalence of sensitive trait is small.

B Computational Details for Measurement Error Models

B.1 The EM Algorithm for the Model of Top-biased Error

We treat (S_i, Z_i, Y_i^*) as (partially) missing data to form the following complete-data likelihood function,

$$\prod_{i=1}^{N} p^{S_i}(1-p)^{1-S_i} g(X_i;\beta)^{T_i Z_i}\{1 - g(X_i;\beta)\}^{T_i(1-Z_i)} \binom{J}{Y_i^*} f(X_i;\gamma)^{Y_i^*}\{1 - f(X_i;\gamma)\}^{J-Y_i^*} \qquad (16)$$

With this much simpler form, we can use the EM algorithm, which consists of a series of weighted regressions, to obtain the ML estimator.

We first derive the E-step. For the latent variable of misreporting, we have,

$$\xi(X_i, T_i, Y_i) \;\; = \;\; \mathbb{E}(S_i \mid X_i, T_i, Y_i) \;\; = \;\; \begin{cases} \frac{p}{p\{1-g(X_i;\beta)f(X_i;\gamma)^J\} + g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\ \frac{p}{p\{1-f(X_i;\gamma)^J\} + f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(0, J) \\ 0 & \text{otherwise} \end{cases}$$

The E-step for the latent variable of truthful response to the sensitive item is given by,

$$
\begin{aligned}
\eta(X_i, 1, Y_i) &= \mathbb{E}(Z_i \mid X_i, T_i = 1, Y_i) \\
&= \begin{cases}
0 & \text{if } i \in \mathcal{J}(1,0) \\
\frac{(1-p)g(X_i;\beta)f(X_i;\gamma)^J + p \cdot g(X_i;\beta)}{p\{1-g(X_i;\beta)f(X_i;\gamma)^J\} + g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\
\frac{g(X_i;\beta)\binom{J}{Y_i-1} f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}}{g(X_i;\beta)\binom{J}{Y_i-1} f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1} + \{1-g(X_i;\beta)\}\binom{J}{Y_i} f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}} & \text{otherwise}
\end{cases}
\end{aligned}
$$

Finally, the E-step for the latent variable representing the response to the control items has several different expressions depending on the values of observed variables. We begin with the control group,

$$
\zeta_J(X_i, 0, Y_i) = \Pr(Y_i^* = J \mid X_i, T_i = 0, Y_i) = \begin{cases}
\frac{f(X_i;\gamma)^J}{p\{1-f(X_i;\gamma)^J\} + f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(0, J) \\
0 & \text{otherwise}
\end{cases}
$$

and for $0 \leq y < J$,

$$
\zeta_y(X_i, 0, Y_i) = \Pr(Y_i^* = y \mid X_i, T_i = 0, Y_i) = \begin{cases}
\frac{p\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{p\{1-f(X_i;\gamma)^J\} + f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(0, J) \\
1 & \text{if } Y_i = y \\
0 & \text{otherwise}
\end{cases}
$$

For the treatment group, we have,

$$
\begin{aligned}
\zeta_J(X_i, 1, Y_i) &= \Pr(Y_i^* = J \mid X_i, T_i = 1, Y_i) \\
&= \begin{cases}
\frac{(1-p)g(X_i;\beta)f(X_i;\gamma)^J + p \cdot f(X_i;\gamma)^J}{p + (1-p)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\
\frac{\{1-g(X_i;\beta)\}f(X_i;\gamma)^J}{\{1-g(X_i;\beta)\}f(X_i;\gamma)^J + g(X_i;\beta) \cdot J \cdot f(X_i;\gamma)^{J-1}\{1-f(X_i;\gamma)\}} & \text{if } i \in \mathcal{J}(1, J) \\
0 & \text{otherwise}
\end{cases} \\
\zeta_0(X_i, 1, Y_i) &= \Pr(Y_i^* = 0 \mid X_i, T_i = 1, Y_i) \\
&= \begin{cases}
\frac{p\{1-f(X_i;\gamma)\}^J}{p + (1-p)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\
1 & \text{if } i \in \mathcal{J}(1, 0) \\
0 & \text{otherwise}
\end{cases}
\end{aligned}
$$

and for $0 < y < J$,

$$
\begin{aligned}
&\zeta_y(X_i, 1, Y_i) = \Pr(Y_i^* = y \mid X_i, T_i = 1, Y_i) \\
&= \begin{cases}
\frac{p\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{p + (1-p)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\
\frac{g(X_i;\beta)\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{g(X_i;\beta)\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y} + \{1-g(X_i;\beta)\}\binom{J}{y+1} f(X_i;\gamma)^{y+1}\{1-f(X_i;\gamma)\}^{J-y-1}} & \text{if } i \in \mathcal{J}(1, y+1) \\
\frac{\{1-g(X_i;\beta)\}\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\{1-g(X_i;\beta)\}\binom{J}{y} f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y} + g(X_i;\beta)\binom{J}{y-1} f(X_i;\gamma)^{y-1}\{1-f(X_i;\gamma)\}^{J-y+1}} & \text{if } i \in \mathcal{J}(1, y) \\
0 & \text{otherwise}
\end{cases}
\end{aligned}
$$

Finally, the Q-function is given by,

$$
\sum_{i=1}^{N} \xi(X_i, T_i, Y_i) \log p + \{1 - \xi(X_i, T_i, Y_i)\} \log(1-p)
$$

$$
\begin{aligned}
&+ \sum_{i=1}^{N} T_i \left[\eta(X_i, 1, Y_i) \log g(X_i;\beta) + \{1-\eta(X_i,1,Y_i)\}\log\{1-g(X_i;\beta)\}\right] \\
&+ \sum_{i=1}^{N} \left\{\sum_{y=1}^{J} y\cdot\zeta_y(X_i,T_i,Y_i)\right\}\log f(X_i;\gamma) + \left\{J - \sum_{y=1}^{J} y\cdot\zeta_y(X_i,T_i,Y_i)\right\}\log\{1-f(X_i;\gamma)\}
\end{aligned}
$$

Thus, the M-step for p is,

$$
p \;=\; \frac{1}{N}\sum_{i=1}^{N}\xi(X_i,T_i,Y_i). \tag{17}
$$

The M-steps for β and γ consist of a series of weighted logistic regressions.

B.2 The EM Algorithm for the Model of Uniform Error

The complete-data likelihood is given by,

$$
\begin{aligned}
&\prod_{i=1}^{n}\{p_1^{S_i}(1-p_1)^{1-S_i}\}^{T_i}\{p_0^{S_i}(1-p_0)^{1-S_i}\}^{1-T_i} \\
&\qquad\times g(X_i;\beta)^{T_iZ_i}\{1-g(X_i;\beta)\}^{T_i(1-Z_i)}\binom{J}{Y_i^*}f(X_i;\gamma)^{Y_i^*}\{1-f(X_i;\gamma)\}^{J-Y_i^*}
\end{aligned} \tag{18}
$$

Then, the EM algorithm, which consists of a series of weighted regressions, is used to obtain the ML estimator.

The E-steps for the latent variables of misreporting and truthful answer to the sensitive item are given by,

$$
\xi(X_i,T_i,Y_i) \;=\; \mathbb{E}(S_i \mid X_i,T_i,Y_i) \;=\; \Pr(S_i=1\mid X_i,T_i,Y_i)
$$

$$
= \begin{cases}
\frac{\frac{p_1}{J+2}}{\frac{p_1}{J+2}+(1-p_1)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i\in\mathcal{J}(1,J+1)\\
\frac{\frac{p_1}{J+2}}{\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\{1-f(X_i;\gamma)\}^J} & \text{if } i\in\mathcal{J}(1,0)\\
\frac{\frac{p_0}{J+1}}{\frac{p_0}{J+1}+(1-p_0)\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}} & \text{if } i\in\mathcal{J}(0,y)\\
\frac{\frac{p_1}{J+2}}{\frac{p_1}{J+2}+(1-p_1)\left[g(X_i;\beta)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}+\{1-g(X_i;\beta)\}\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}\right]} & \text{otherwise}
\end{cases}
$$

$$
\eta(X_i,T_i=1,Y_i) \;=\; \mathbb{E}(Z_i \mid X_i,T_i=1,Y_i)
$$

$$
= \begin{cases}
\frac{\frac{p_1}{J+2}g(X_i;\beta)+(1-p_1)g(X_i;\beta)f(X_i;\gamma)^J}{(1-p_1)g(X_i;\beta)f(X_i;\gamma)^J+\frac{p_1}{J+2}} & \text{if } i\in\mathcal{J}(1,J+1)\\
\frac{\frac{p_1}{J+2}g(X_i;\beta)}{(1-p_1)\{1-g(X_i;\beta)\}\{1-f(X_i;\gamma)\}^J+\frac{p_1}{J+2}} & \text{if } i\in\mathcal{J}(1,0)\\
\frac{\left[\frac{p_1}{J+2}+(1-p_1)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}\right]g(X_i;\beta)}{\frac{p_1}{J+2}+(1-p_1)\left[g(X_i;\beta)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}+\{1-g(X_i;\beta)\}\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}\right]} & \text{otherwise}
\end{cases}
$$

For the latent variable of response to the control items, we obtain the E-steps separately for different sets of observations. For the control group, we have

$$
\zeta_y(X_i,0,Y_i) \;=\; \Pr(Y_i^*=y\mid X_i,T_i=0,Y_i) \;=\; \begin{cases}
\frac{\left[\frac{p_0}{J+1}+(1-p_0)\right]\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_0}{J+1}+(1-p_0)\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}} & \text{if } i\in\ \mathcal{J}(0,y)\\
\frac{\frac{p_0}{J+1}\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_0}{J+1}+(1-p_0)\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}} & \text{otherwise}
\end{cases}
$$

where $y = 0, 1, \ldots, J$. For the treatment group, the E-step is more complex,

$$\zeta_J(X_i, 1, Y_i) \;=\; \Pr(Y_i^* = J \mid X_i, T_i = 1, Y_i)$$

$$= \begin{cases} \frac{\left\{(1-p_1)g(X_i;\beta)+\frac{p_1}{J+2}\right\}f(X_i;\gamma)^J}{\frac{p_1}{J+2}+(1-p_1)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\ \frac{\left[(1-p_1)\{1-g(X_i;\beta)\}+\frac{p_1}{J+2}\right]f(X_i;\gamma)^J}{(1-p_1)[\{1-g(X_i;\beta)\}f(X_i;\gamma)^J+g(X_i;\beta)Jf(X_i;\gamma)^{J-1}\{1-f(X_i;\gamma)\}]+\frac{p_1}{J+2}} & \text{if } i \in \mathcal{J}(1, J) \\ \frac{\frac{p_1}{J+2}f(X_i;\gamma)^J}{\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\{1-f(X_i;\gamma)\}^J} & \text{if } i \in \mathcal{J}(1, 0) \\ \frac{\frac{p_1}{J+2}f(X_i;\gamma)^J}{(1-p_1)\left[\{1-g(X_i;\beta)\}\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}+g(X_i;\beta)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}\right]+\frac{p_1}{J+2}} & \text{otherwise} \end{cases}$$

$$\zeta_0(X_i, 1, Y_i) \;=\; \Pr(Y_i^* = 0 \mid X_i, T_i = 1, Y_i)$$

$$= \begin{cases} \frac{\frac{p_1}{J+2}\{1-f(X_i;\gamma)\}^J}{\frac{p_1}{J+2}+(1-p_1)g(X_i;\beta)f(X_i;\gamma)^J} & \text{if } i \in \mathcal{J}(1, J+1) \\ \frac{\left[\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\right]\{1-f(X_i;\gamma)\}^J}{\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\{1-f(X_i;\gamma)\}^J} & \text{if } i \in \mathcal{J}(1, 0) \\ \frac{\left[(1-p_1)\{1-g(X_i;\beta)\}+\frac{p_1}{J+2}\right]\{1-f(X_i;\gamma)\}^J}{(1-p_1)\left[\{1-g(X_i;\beta)\}\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}+g(X_i;\beta)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}\right]+\frac{p_1}{J+2}} & \text{otherwise} \end{cases}$$

and for $0 < y < J$, we have,

$$\zeta_y(X_i, 1, Y_i) \;=\; \Pr(Y_i^* = y \mid X_i, T_i = 1, Y_i)$$

$$= \begin{cases} \frac{\left\{\frac{p_1}{J+2}+(1-p_1)g(X_i;\beta)\right\}\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_1}{J+2}+(1-p_1)\left[g(X_i;\beta)\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}+\{1-g(X_i;\beta)\}\binom{J}{y+1}f(X_i;\gamma)^{y+1}\{1-f(X_i;\gamma)\}^{J-y-1}\right]} & \text{if } i \in \mathcal{J}(1, y+1) \\ \frac{\left[\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\right]\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_1}{J+2}+(1-p_1)\left[g(X_i;\beta)\binom{J}{y-1}f(X_i;\gamma)^{y-1}\{1-f(X_i;\gamma)\}^{J-y+1}+\{1-g(X_i;\beta)\}\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}\right]} & \text{if } i \in \mathcal{J}(1, y) \\ \frac{\frac{p_1}{J+2}\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_1}{J+2}+(1-p_1)\{1-g(X_i;\beta)\}\{1-f(X_i;\gamma)\}^J} & \text{if } i \in \mathcal{J}(1, 0) \\ \frac{\frac{p_1}{J+2}\binom{J}{y}f(X_i;\gamma)^y\{1-f(X_i;\gamma)\}^{J-y}}{\frac{p_1}{J+2}+(1-p_1)\left[\{1-g(X_i;\beta)\}\binom{J}{Y_i}f(X_i;\gamma)^{Y_i}\{1-f(X_i;\gamma)\}^{J-Y_i}+g(X_i;\beta)\binom{J}{Y_i-1}f(X_i;\gamma)^{Y_i-1}\{1-f(X_i;\gamma)\}^{J-Y_i+1}\right]} & \text{otherwise} \end{cases}$$

Finally, the Q-function is given by,

$$\begin{aligned} &\sum_{i=1}^{n} \mathbb{E}(S_i \mid X_i, T_i = 1, Y_i) \log p_1 + \{1 - \mathbb{E}(S_i \mid X_i, T_i = 1, Y_i)\} \log(1 - p_1) \\ &+ \mathbb{E}(S_i \mid X_i, T_i = 0, Y_i) \log p_0 + \{1 - \mathbb{E}(S_i \mid X_i, T_i = 0, Y_i)\} \log(1 - p_0) \\ &\mathbb{E}(Z_i \mid X_i, T_i = 1, Y_i) \log g(X_i; \beta) + \{1 - \mathbb{E}(Z_i \mid X_i, T_i = 1, Y_i)\} \log\{1 - g(X_i; \beta)\} \\ &+ \mathbb{E}(Y_i^* \mid X_i, T_i, Y_i) \log f(X_i; \gamma) + \{J - \mathbb{E}(Y_i^* \mid X_i, T_i, Y_i)\} \log\{1 - f(X_i; \gamma)\} \end{aligned} \tag{19}$$

Hence, the M-steps for p_0 and p_1 are immediate. The M-steps for β and γ consist of a series of weighted logistic regressions.

B.3 Details of the Robust Maximum Likelihood Multivariate Regression Estimator

We focus on the logistic regression model whose log-likelihood function is given,

$$-\sum_{i=1}^{N} \Big[J \log\{1 + \exp(X_i^\top \gamma)\} + \log\{1 + \exp(X_i^\top \beta)\} \Big] + \sum_{i \in \mathcal{J}(1, J+1)} \Big(X_i^\top \beta + J X_i^\top \gamma \Big) +$$

$$\sum_{y=0}^{J}\sum_{i\in\mathcal{J}(0,y)}\left[yX_i^\top\gamma+\log\{1+\exp(X_i^\top\beta)\}\right]+$$
$$\sum_{y=1}^{J}\sum_{i\in\mathcal{J}(1,y)}\left[(y-1)X_i^\top\gamma+\log\left\{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)\right\}\right]+\text{constant} \tag{20}$$

Let $\mathcal{L}_i(\beta,\gamma;X_i,Y_i)$ represent the log-likelihood function for observation i. Then, the first order condition for each observation is given by,

$$\begin{aligned}&\frac{\partial}{\partial\beta}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\ =\;&\left[-\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}+\mathbf{1}\{i\in\mathcal{J}(1,J+1)\}+\right.\\ &\left.\sum_{y=0}^{J}\mathbf{1}\{i\in\mathcal{J}(0,y)\}\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}+\sum_{y=1}^{J}\mathbf{1}\{i\in\mathcal{J}(1,y)\}\frac{\binom{J}{y-1}\exp(X_i^\top\beta)}{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)}\right]X_i\end{aligned} \tag{21}$$

$$\begin{aligned}&\frac{\partial}{\partial\gamma}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\ =\;&\left[-\frac{J\exp(X_i^\top\gamma)}{1+\exp(X_i^\top\gamma)}+J\mathbf{1}\{i\in\mathcal{J}(1,J+1)\}+\right.\\ &\left.\sum_{y=0}^{J}y\mathbf{1}\{i\in\mathcal{J}(0,y)\}+\sum_{y=1}^{J}\mathbf{1}\{i\in\mathcal{J}(1,y)\}\left((y-1)+\frac{\binom{J}{y}\exp(X_i^\top\gamma)}{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)}\right)\right]X_i\end{aligned} \tag{22}$$

The sample analogue of the moment condition given in equation (13) can be written as,

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{M}_i(\beta;X_i,Y_i)\;=\;\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}-\hat{\tau}\right)\;=\;0 \tag{23}$$

where $\hat{\tau}$ is the difference-in-means estimator. We can also express this condition as

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{M}_i(\beta;X_i,Y_i)\;=\;\frac{1}{N}\sum_{i=1}^{N}\left[T_i\left(\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}-\frac{N}{N_1}Y_i\right)+(1-T_i)\left(\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}+\frac{N}{N_0}Y_i\right)\right], \tag{24}$$

in order to account for the correlation between this moment and the score function.

Putting together all these moment conditions, the efficient GMM estimator is given by,

$$(\hat{\beta}_{\text{GMM}},\hat{\gamma}_{\text{GMM}})\;=\;\arg\min_{(\beta,\gamma)}\mathcal{G}(\beta,\gamma)^\top\mathcal{W}(\beta,\gamma)^{-1}\mathcal{G}(\beta,\gamma) \tag{25}$$

where

$$\mathcal{G}(\beta,\gamma)\;=\;\frac{1}{N}\sum_{i=1}^{N}\mathcal{G}_i(\beta,\gamma)\;=\;\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\beta}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)^\top\;\frac{\partial}{\partial\gamma}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)^\top\;\mathcal{M}_i(\beta;X_i,Y_i)^\top\right]^\top \tag{26}$$

$$\mathcal{W}(\beta,\gamma)\;=\;\frac{1}{N}\sum_{i=1}^{N}\mathcal{G}_i(\beta,\gamma)\mathcal{G}_i(\beta,\gamma)^\top. \tag{27}$$

The asymptotic distribution of this estimator is given by:

$$\sqrt{N}\left(\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix}-\begin{pmatrix}\beta\\ \gamma\end{pmatrix}\right) \rightsquigarrow \mathcal{N}\left(0,\ \left[\left(\mathbb{E}\frac{\partial \mathcal{G}_i(\beta,\gamma)}{\partial(\beta^\top\ \gamma^\top)^\top}\right)^\top \Omega(\beta,\gamma)^{-1}\mathbb{E}\frac{\partial \mathcal{G}_i(\beta,\gamma)}{\partial(\beta^\top\ \gamma^\top)^\top}\right]^{-1}\right) \tag{28}$$

where

$$\mathbb{E}\frac{\partial \mathcal{G}_i(\beta,\gamma)}{\partial(\beta^\top\ \gamma^\top)^\top} = \mathbb{E}\begin{pmatrix}\frac{\partial^2}{\partial\beta\partial\beta^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i) & \frac{\partial^2}{\partial\beta\partial\gamma^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\ \frac{\partial^2}{\partial\gamma\partial\beta^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i) & \frac{\partial^2}{\partial\gamma\partial\gamma^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\ \frac{\partial}{\partial\beta^\top}\mathcal{M}_i(\beta;X_i,Y_i) & 0\end{pmatrix} \tag{29}$$

and

$$\Omega(\beta,\gamma) = \mathbb{E}\left[\left(\frac{\partial \mathcal{G}_i(\beta,\gamma)}{\partial(\beta^\top\ \gamma^\top)^\top}\right)\left(\frac{\partial \mathcal{G}_i(\beta,\gamma)}{\partial(\beta^\top\ \gamma^\top)^\top}\right)^\top\right] \tag{30}$$

Note that the second derivatives are given by,

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta\partial\beta^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\
=\ &\left[-\frac{\exp(X_i^\top\beta)}{\{1+\exp(X_i^\top\beta)\}^2}+\sum_{y=0}^{J}\mathbf{1}\{i\in\mathcal{J}(0,y)\}\frac{\exp(X_i^\top\beta)}{\{1+\exp(X_i^\top\beta)\}^2}+\right.\\
&\left.\sum_{y=1}^{J}\mathbf{1}\{i\in\mathcal{J}(1,y)\}\frac{\exp\left\{\binom{J}{y-1}\binom{J}{y}X_i^\top(\gamma+\beta)\right\}}{\left\{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)\right\}^2}\right]X_iX_i^\top
\end{aligned} \tag{31}$$

$$\begin{aligned}
&\frac{\partial^2}{\partial\gamma\partial\gamma^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\
=\ &\left[-\frac{J\exp(X_i^\top\gamma)}{\{1+\exp(X_i^\top\gamma)\}^2}+\sum_{y=1}^{J}\mathbf{1}\{i\in\mathcal{J}(1,y)\}\frac{\exp\left\{\binom{J}{y-1}\binom{J}{y}X_i^\top(\gamma+\beta)\right\}}{\left\{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)\right\}^2}\right]X_iX_i^\top
\end{aligned} \tag{32}$$

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta\partial\gamma^\top}\mathcal{L}_i(\beta,\gamma;X_i,Y_i)\\
=\ &-\left[\sum_{y=1}^{J}\mathbf{1}\{i\in\mathcal{J}(1,y)\}\frac{\exp\left\{\binom{J}{y-1}\binom{J}{y}X_i^\top(\gamma+\beta)\right\}}{\left\{\binom{J}{y-1}\exp(X_i^\top\beta)+\binom{J}{y}\exp(X_i^\top\gamma)\right\}^2}\right]X_iX_i^\top
\end{aligned} \tag{33}$$

C Additional Simulation Results

In this section, we present additional simulation results that build on the results shown in Section 3.

C.1 Addressing Response Error

The following Figure 5 illustrates the properties of the ML and novel measurement error models introduced in this paper when we increase the underlying prevalence of the sensitive trait. The corresponding figure in the paper is Figure 2. Increasing the underlying prevalence tends to improve the performance of all the models, although the ML and constrained ML models remain positively biased. The constrained models improve inference for the prevalence of the sensitive trait regardless of the measurement error mechanism. However, it is not possible to improve inference for the coefficients under measurement error without assuming the error mechanism and using the corresponding model.

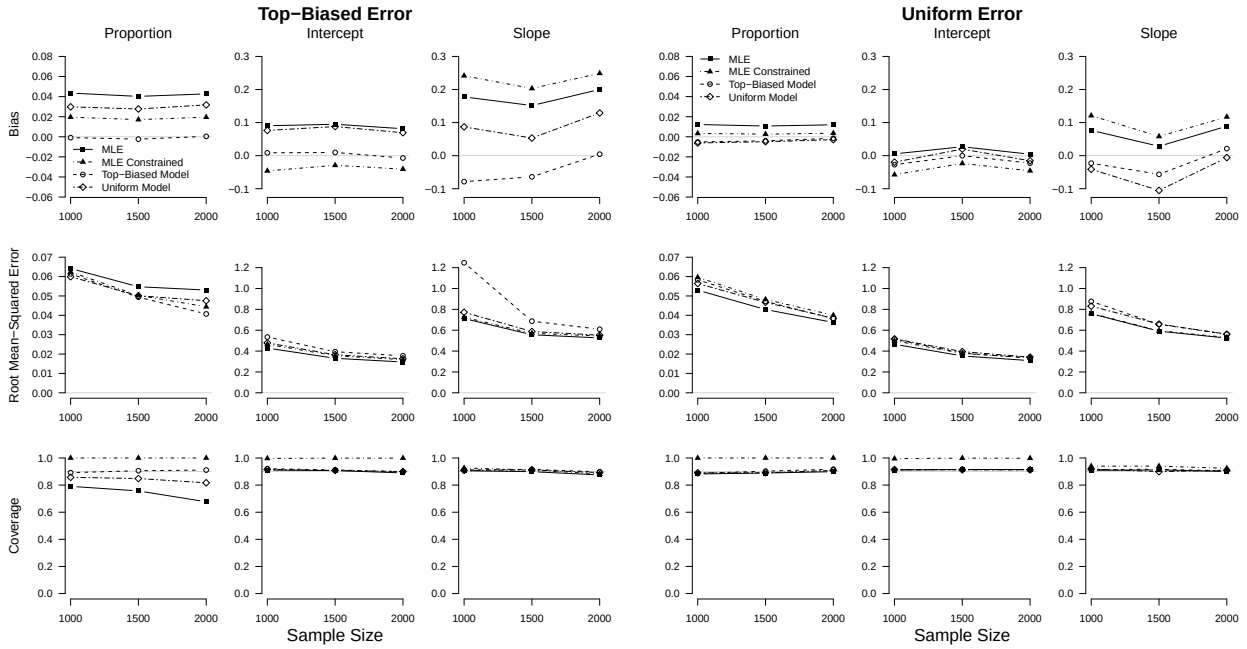

Figure 5: Robustness of the Constrained and Measurement Error Maximum Likelihood Estimators in the Presence of Response Errors when the Propensity of Sensitive Trait is High. We consider four estimators of the prevalence of sensitive trait and slope and intercept regression coefficients: the standard maximum likelihood (ML) estimator (solid square with solid line), the constrained ML estimator (solid triangle with dot-dash line), the ML estimators adjusting for top-biased response error (open circle with dashed lines) and uniform response errors (open diamond with dot-long-dash line). The result shows that both the constrained MLE estimator and the models adjusting for response error are an improvement over the performance of the MLE estimator.

C.2 Results for $J = 4$

Figure 6 replicates Figure 1 with the number of control items increased from $J = 3$ to 4. The substantive conclusions remain the same.

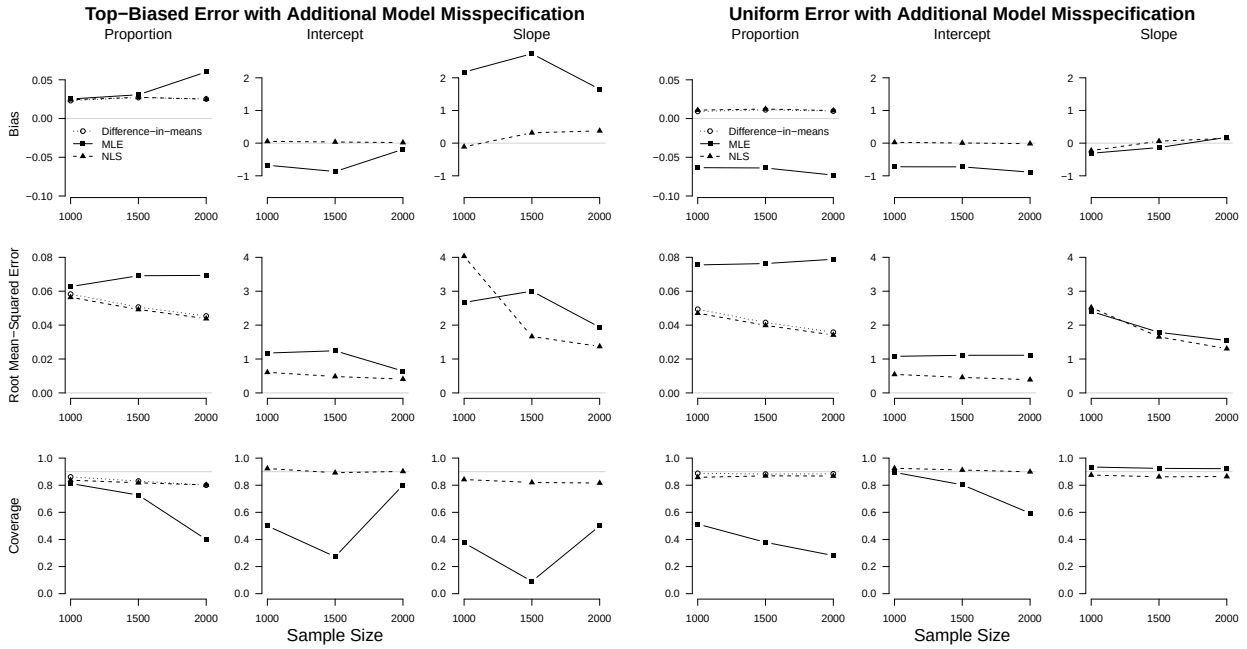

Figure 6: Robustness of the Nonlinear Least Squares Regression Estimator in the Presence of Several Model Misspecifications Considered in Ahlquist (2017) for $J = 4$. We consider the three estimators of the prevalence of sensitive trait: the difference-in-means estimator (open circle with dotted line), the maximum likelihood (ML) regression estimator (solid square with solid line), and the nonlinear least squares (NLS) estimator (solid triangle with dashed line). The result shows that the NLS regression estimator is as robust as the difference-in-means estimator.

Figure 7 replicates Figure 2 with the number of control items creased from $J = 3$ to 4. This somewhat increases the bias of the estimated slope coefficients. However, the substantive conclusions are unchanged. Specifically, the constrained models help the proportion but not the covariates, while the measurement error models are needed to improve inference for the covariates. Again, the standard MLE is more robust to uniform measurement error than to top-biased error.

Figure 8 replicates Figure 5 with the number of control items increased from $J = 3$ to 4. The substantive conclusions remain the same. However, the uniform error model results in a slight underestimate of the proportion.

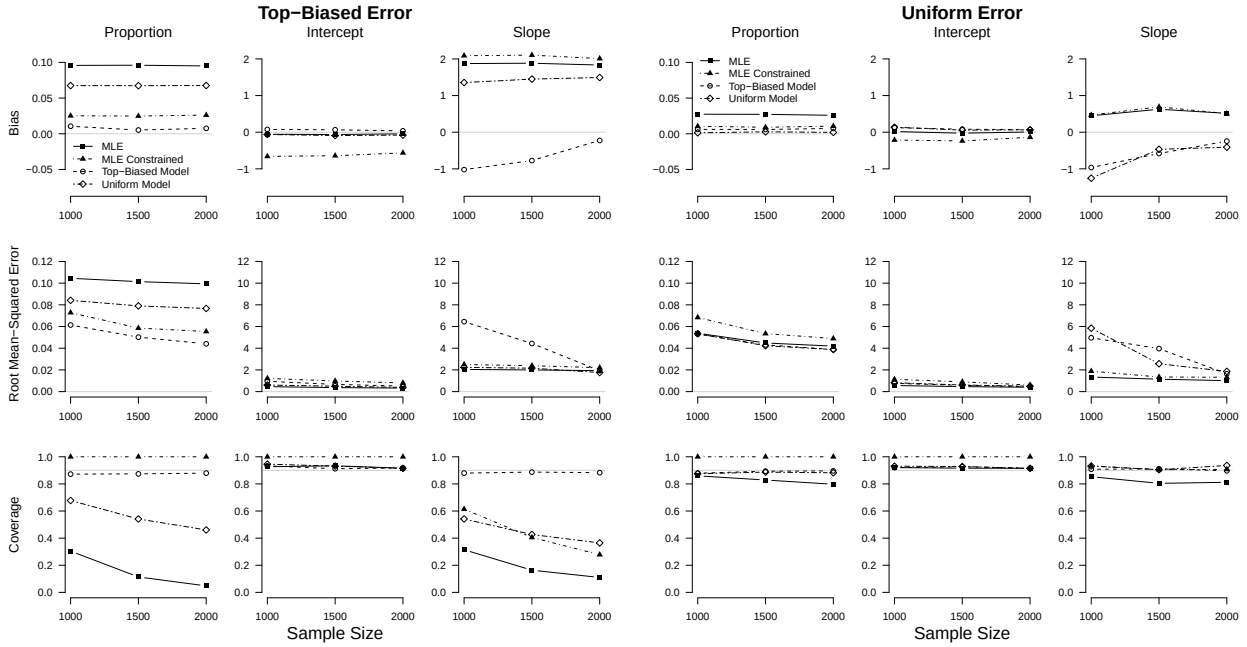

Figure 7: Robustness of the Constrained Maximum Likelihood and Measurement Error Maximum Likelihood Estimators in the Presence of Response Errors when the Propensity of Sensitive Trait is Low for $J = 4$. We consider the four estimators of the prevalence of sensitive trait and slope and intercept regression coefficients: the standard maximum likelihood (ML) estimator (solid square with solid line), the constrained ML estimator (solid triangle with dot-dash line), the ML estimators adjusting for top-biased response error (open square with dashed lines) and uniform response errors (open diamond with dot-long-dash line). The result shows that both the constrained MLE estimator and the models adjusting for response error are an improvement over the performance of the MLE estimator.

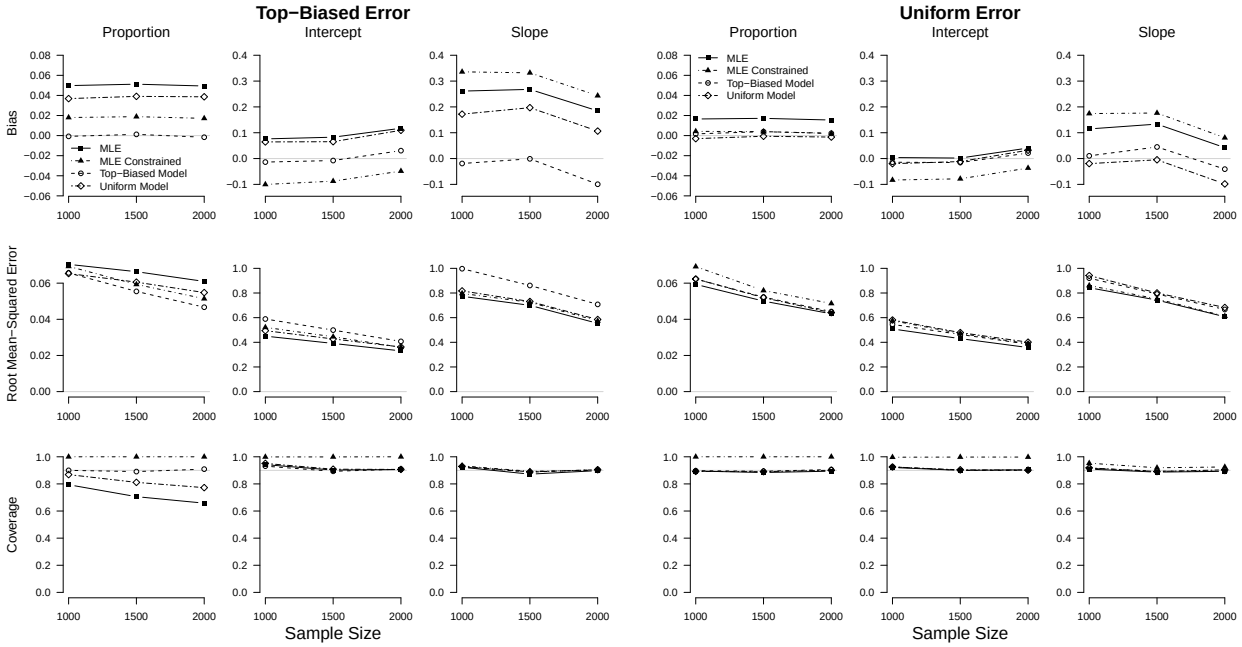

Figure 8: Robustness of the Constrained Maximum Likelihood and Measurement Error Maximum Likelihood Estimators in the Presence of Response Errors when the Propensity of Sensitive Trait is High for $J = 3$. We consider the four estimators of the prevalence of sensitive trait and slope and intercept regression coefficients: the standard maximum likelihood (ML) estimator (solid square with solid line), the constrained ML estimator (solid triangle with dot-dash line), the ML estimators adjusting for top-biased response error (open square with dashed lines) and uniform response errors (open diamond with dot-long-dash line). The result shows that both the constrained MLE estimator and the models adjusting for response error are an improvement over the performance of the MLE estimator.