Item Count Technique Estimators under Respondent Error

John S. Ahlquist*

October 28, 2016

Abstract

In recent years we have seen the introduction of new tools for the analysis of survey list experiments, most notably the Item Count Technique (ICT-MLE) regression model (Imai, 2011). This estimator promises to extract more from the data–and more efficiently–than traditional difference-in-means analysis but it leans heavily on assumptions about responses at the extremes (choosing no or all items on the list). I argue that the assumptions required to identify and estimate the ICT-MLE are unlikely to hold in many applied settings. I document that such problems arise in practice and then report the results of Monte Carlo experiments examining the sensitivity of two estimators to various types of respondent error. I find that both the difference in means estimator and the the ICT-MLE are sensitive to measurement error, but the problems are much more severe for the ICT-MLE. Respondent error directed disproportionately toward the extreme responses is most pernicious for both estimators. This bias in the ICT-MLE becomes more extreme as the underlying population prevalence of the sensitive item declines. I provide guidance for extra care when considering the use of the ICT-MLE as well as suggestions for further research.

*associate professor, School of Global Policy and Strategy, UC San Diego. jahlquist@ucsd.edu. Versions of this paper were presented at the 2014 PolMeth and Midwest Political Science Association meetings as well the UW-Madison Models and Data group and colloquia at UC San Diego and the University of Washington Center for Statistics and the Social Sciences. I thank Graeme Blair, Scott Gehlbach, Kosuke Imai, Simon Jackman, Kelly Matush, Tom Pepinsky, Molly Roberts, Mike Ward, Yiqing Xu and, especially, Alex Tahk, for helpful conversations.

Social scientists are often interested in people's thoughts and behavior around sensitive issues, such as racial attitudes, sexual behaviors, participation in illegal or undesirable activities, or simply eliciting an honest report of voter turnout. Directly asking people about such topics is unlikely to generate truthful responses so researchers have developed several methods of indirect questioning.[1] Survey list experiments, originally developed decades ago, are one potentially useful tool for indirectly asking respondents to reveal their thoughts on sensitive topics.

Traditional list experiments achieve this by randomly dividing respondents into two groups. Respondents in the baseline (or control) group are presented with a list of innocuous items and asked *how many* (as opposed to which) items pertain to them. Respondents in the treatment group are presented with the same list of items along with one additional item describing the sensitive attribute of interest. Asking respondents to report a number protects their individual privacy around the sensitive item—unless those in the treatment group select the maximum or minimum value. Assuming randomization worked appropriately, the only difference, on average, between the treatment and control groups is the number of items on the lists they see. The difference in the average number of items reported by members of the treatment and control groups is therefore an estimate of the prevalence of the sensitive item in the larger population from which the sample was drawn. There have been numerous applications of list experiments in recent years. Restricting attention to just studies of voter fraud and election irregularities, list experiments have been deployed in Lebanon (Corstange, 2009, 2012), Nicaragua (Gonzalez-Ocantos et al., 2012), Russia (Frye, Reuter and Szakonyi, 2014), and the United States (Ahlquist, Mayer and Jackman, 2014). Kiewiet de Jonge and Nickerson (2013) report numerous other applications from comparative and American politics.[2]

[1] See Blair, Imai and Lyall (2014); Rosenfeld, Imai and Shapiro (2015) for discussions and comparisons across several methods of indirect questioning.

[2] See also Gingerich (2010), Glynn (2013)

With the advent of cheap and reliable Internet survey panels list experiments are easier and cheaper to deploy than ever. Along with increased interest in list experiments have come new design procedures (Glynn, 2013) and statistical estimators for list experiment data, most notably the Item Count Technique (ICT) regression models (Blair and Imai, 2012; Imai, 2011). These estimators purport to tell us much more about the relationships between covariates and the sensitive behavior than traditional difference-in-means analysis (DiM).

Imai proposes a non-linear least squares and a fully specified maximum likelihood estimator. The last of these is the key statistical innovation that has justifiably received the most attention. This paper interrogates one of the key assumptions of the ICT maximum likelihood estimator (ICT-MLE). Specifically, the ICT-MLE leans heavily on information contained in responses at the extremes of the treatment group (admitting to 0 or all items). Consequently the estimator requires strong assumptions about the truthfulness of respondents' answers (Imai's *no liars* assumption). I argue that in actual applied settings responses in this part of the distribution are particularly likely to result from error. The ICT-MLE will be relatively sensitive to measurement or respondent error, especially when it appears in the extremes.

Blair and Imai (2012); Chaudhuri and Christofides (2007); Glynn (2013) do important work developing diagnostic tests and modeling extensions for *strategic* misrepresentation (so-called floor and ceiling effects). Kuha and Jackson (2014) extend and improve the ICT-MLE algorithm and variance estimation. But there is no work systematically examining the robustness of these estimators under measurement error. In this paper I recapitulate how assumptions regarding respondent accuracy are critical to the performance of the ICT estimators and show that the assumptions required for unbiasedness in the DiM estimator are weaker than those for ICT-MLE. I show that the ICT-MLE estimator is particularly sensitive to violations of the no liars assumption. I then demonstrate that respondent error

and estimator bias are more than just a hypothetical concerns; they can arise at non-trivial levels in real applications that pass existing statistical tests.

The core of the paper presents results from a series of Monte Carlo experiments comparing the performance of ICT-MLE to simple difference-in-means analysis under several hypothetical patterns of respondent error. The Monte Carlo results lead to four conclusions: 1) the ICT-MLE is biased even under no respondent error (contrary to earlier results); 2) While respondent error induces bias in both the difference-in-means and ICT-ML estimators, the ICT-ML estimator is far more sensitive to measurement error, especially when this error manifests disproportionately as extreme-valued responses; 3) The bias induced by respondent error becomes more extreme as the underlying prevalence of the sensitive item decreases; 4) Given the bias in the ICT-ML estimator, a simple Hausman-style specification test does *not* perform adequately. I conclude by offering some conjectures about when and where respondent error is most likely to arise as well as possible strategies for mitigating these problems.

1 Extracting information from indirect questions

List experiments are an indirect questioning method designed to measure the prevalence of a particular attribute or behavior in some population. As running example, consider the voter impersonation list experiment used in Ahlquist, Mayer and Jackman (2014). Figure 1 displays the question (and the YouGov user interface) for the treatment group.

List experiments have costs relative to direct questioning: they are harder to administer, they are a less efficient use of the sample, and they may be confusing or off-putting to some respondents. A researcher would therefore resort to a list experiment only when she has reason to believe that:

1. For the sensitive topic there are respondents who do not want their answers to be

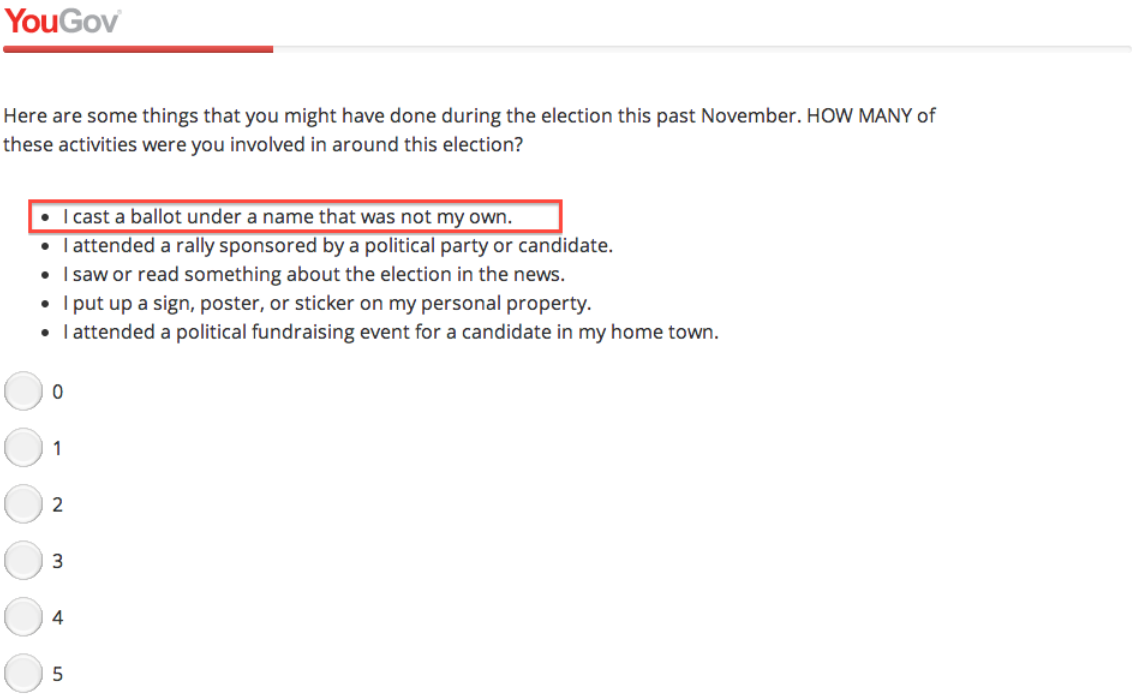

Figure 1: An example of the user interface facing respondents to the YouGov survey employed in Ahlquist, Mayer and Jackman (2014). The sensitive item is highlighted here; actual survey respondents would not see the red box. Respondents in the baseline condition would see only four items with the sensitive item omitted. The ordering of items in the list was randomized.

traceable to them individually, even if survey data are reported as anonymous and even if the only person with any knowledge of the individual response is a survey enumerator.

2. For the sensitive topic, at least some of the reticent respondents do harbor some latent desire to answer truthfully and would do so given additional identity protection.[3]

List experiments do not automatically solve the problems that motivate their use, however. Respondents in two situations remain compromised: (i) those in the treatment group who would answer affirmatively to all of the baseline items and the sensitive item and (ii) those in the treatment group who answer negatively to all the baseline items and the sen-

[3]See a useful set of reflections on list experiments in Gelman (2014).

sitive item. These respondents are still forced to choose between either truthfully revealing their status or misrepresenting their answers. To the extent these situations are present and respondents dissemble the list experiment is said to exhibit ceiling and floor effects, respectively.

Survey design best practices recognize this. Kuklinski, Cobb and Gilens (1997) advise that an appropriately designed list experiment will seek to minimize the number of respondents put in this exact situation. Glynn (2013) urges applied researchers to identify negatively correlated control items with non-trivial population rates to achieve a list that avoids ceiling and floor effects while also minimizing the variance of the difference-in-means estimator.

At the same time a new set of statistical tools have emerged for analyzing data from list experiments, claiming to extract more information more efficiently than simple difference-in-means tools. Existing work, however, fails to consider the consequences of measurement error for these newer estimators.

1.1 Notation

We introduce some notation and definitions that closely follow Blair and Imai (2012); Imai (2011).

We assume a random sample of N respondents from some population. Sample members are indexed by i. Respondents confront a standard design list experiment in which there are J control items. The indicator T_i denotes whether i sees the list with just J items ($T_i = 0$) or sees the list with J control items and the additional sensitive item. Let $C_{i1}(t), \ldots, C_{iJ}(t)$ denote i's latent response to each control item as a function of whether the respondent sees the J-item ($T_i = 0$) or $(J+1)$-item list ($T_i = 1$); $C_{ij}(t) = 1$ implies an affirmative latent response to control item j under treatment condition t. Let $Z_i(1)$ denote i's latent response to the sensitive item under the treatment condition and let Z_i^* denote i's truthful response

to the sensitive item. Potential outcomes, $Y_i(t)$, are defined as

$$Y_i(1) = Z_i(1) + \sum_{j=1}^{J} C_{ij}(1) \quad (1)$$

$$Y_i(0) = \sum_{j=1}^{J} C_{ij}(0) \quad (2)$$

Observed data are simply $Y_i(T_i)$.

The quantity of ultimate interest is the population prevalence of the sensitive item, i.e., $\Pr(Z_i^* = 1) \equiv \pi_{Z^*}$. Secondary quantities of interest may include parameters, θ, that describe $\Pr(Z_i^* = 1 \mid X_i; \theta)$.

To identify the model below, Imai (2011) introduces three assumptions:

- *Randomization:* $T_i \perp \{Z_i(1), C_{ij}(1), C_{ij}(0)\} \forall i$
- *No design effects:* $\sum_{j=1}^{J} C_{ij}(0) = \sum_{j=1}^{J} C_{ij}(1) \forall i$
- *No liars:* $Z_i^* = Z_i(1) \forall i$

Note that the *no design effects* and *no liars* assumptions jointly imply the assumption of no measurement error correlated with treatment *at the individual level* and that any measurement error that does exist occurs *only* among the control items.

1.2 Difference in means estimator

The difference-in-means estimator of π_{Z^*}, henceforth DiM, is simply

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N - N_1} \sum_{i=1}^{N} (1 - T_i) Y_i \quad (3)$$

where N_1 is the number of respondents in the treatment condition.

It is straight forward to show that OLS regression of Y on T is equivalent to DiM estimation. Importantly, we can also show that the DiM estimator is unbiased under weaker conditions than those stated above. If we allow for measurement error, e_i, such that $Y_i = Y_i^* + e_i$, then the DiM estimator becomes

$$\begin{aligned} Y_i^* &= \alpha + \tau T_i + \varepsilon_i \\ \varepsilon_i &= e_i + \epsilon_i \end{aligned} \tag{4}$$

where ϵ represents random sampling variation in Y_i. From this expression, it is clear to see that $\hat{\tau}$ is an unbiased estimator of π_{Z^*} so long as $\text{Cov}(T_i, \varepsilon_i) = 0$. The random assignment of T_i achieves this so long as measurement error is uncorrelated with T_i. But note that measurement error can occur anywhere in the response space so long as it is uncorrelated with treatment *on average*. In other words, the DiM estimator only requires two sums; measurement error can occur anywhere in the sum so long as, on average, the error is expected to even out across treatment and control groups. Furthermore, the amount of bias in $\hat{\tau}$ is directly determined by the magnitude of e_i and the extent to which it is correlated with T_i.

1.3 Identifying joint proportions

The DiM estimator relies on relatively weak assumptions. Glynn (2013) invokes (by other names) the stronger *no design effects* and *no liars* assumptions in order to characterize the joint distribution of $(Y_i(0), Z_i^*)$, thereby generating estimates of the population proportion who would answer affirmatively to the sensitive item and exactly y of the control items. It is then a short jump to including covariates.

To see this intuitively, consider a list experiment with $J = 4$ baseline items, as in Table 1.[4]

[4]See Glynn (2013:appendix D) for a more general formal derivation.

Define $\mathcal{K}(y, z^*)$ as the set of individuals with values $(Y_i(0), Z_i^*)$, i.e. the set that would respond affirmatively to y baseline items (under the no-treatment condition) and with true response of z^* for the sensitive item. Let π_{y1} be the population proportion of people who would answer yes to y control items and the sensitive item. Under the maintained identification assumptions we know that all respondents who answer "0" in the treatment condition–the first cell in the third column of the table–are certainly in $\mathcal{K}(0, 0)$ and all who answer "5" are in $\mathcal{K}(4, 1)$. The remainder of the table describes the other combinations.

Table 1: Respondent types identified under the *no design effect* and *no liars* assumptions for a $J = 4$ list experiment.

y_i	Baseline	Treatment
0	$\mathcal{K}(0,0) \cup \mathcal{K}(0,1)$	$\mathcal{K}(0,0)$
1	$\mathcal{K}(1,0) \cup \mathcal{K}(1,1)$	$\mathcal{K}(1,0) \cup \mathcal{K}(0,1)$
2	$\mathcal{K}(2,0) \cup \mathcal{K}(2,1)$	$\mathcal{K}(2,0) \cup \mathcal{K}(1,1)$
3	$\mathcal{K}(3,0) \cup \mathcal{K}(3,1)$	$\mathcal{K}(3,0) \cup \mathcal{K}(2,1)$
4	$\mathcal{K}(4,0) \cup \mathcal{K}(4,1)$	$\mathcal{K}(4,0) \cup \mathcal{K}(3,1)$
5	$\varnothing$	$\mathcal{K}(4,1)$

From here we can characterize other quantities, for example $\hat{\pi}_{21}$. By the identification assumptions, the baseline respondents answering "3" or higher are all the baseline respondents in $\mathcal{K}(3,0) \cup \mathcal{K}(3,1) \cup \mathcal{K}(4,0) \cup \mathcal{K}(4,1)$. Similarly everyone who answers "3" or higher in the treatment condition are the treatment respondents in $\mathcal{K}(2,1) \cup \mathcal{K}(3,0) \cup \mathcal{K}(3,1) \cup \mathcal{K}(4,0) \cup \mathcal{K}(4,1)$. The disjointness of all these $\mathcal{K}(y, z*)$ sets implies that $|\{y_i : y_i \geq 3, T_i = 0\}|/(N-N_1)$ is an unbiased estimator of $(\pi_{30} + \pi_{31} + \pi_{40} + \pi_{41})$. Similarly $|\{y_i : y_i \geq 3, T_i = 1\}|/N_1$ is an unbiased estimator of $(\pi_{21} + \pi_{30} + \pi_{31} + \pi_{40} + \pi_{41})$. Thus an unbiased estimator of $\hat{\pi}_{21}$ is

$$\hat{\pi}_{21} = |\{y_i : y_i \geq 3, T_i = 1\}|/N_1 - |\{y_i : y_i \geq 3, T_i = 0\}|/(N - N_1)$$

Clearly this exercise can be repeated to estimate any of the π_{y1} quantities. Summing the $\hat{\pi}_{y1}$ yields the DiM estimator in Equation 14 (see the appendix for a derivation).

1.4 The ICT-ML model

Imai (2011) develops a maximum likelihood estimator for the joint distribution $(Y_i(0), Z_i(t))$. The ICT-MLE is explicitly designed to generate better descriptions of the population heterogeneity of the sensitive item by efficiently including covariates in the model. The ICT-MLE also generates individual-level predicted probabilities that a respondent possesses the sensitive attribute.

Define $\mathcal{J}(t, y)$ as the set of respondents with values $(T_i, Y_i) = (t, y)$. Let $g(\mathbf{x}, \delta) = \Pr(Z_i(1) = 1 \mid \mathbf{X}_i = \mathbf{x})$ and $h_z(y; \mathbf{x}, \psi_{\mathbf{z}}) = \Pr(Y_i(0) = y \mid \mathbf{X}_i = \mathbf{x}, Z_i(1) = z)$, where $\mathbf{x}$ represents a vector of covariates and $\delta, \psi_{\mathbf{z}}$ are parameter vectors to be estimated. The observed data likelihood can now be stated as

$$\begin{aligned}
\mathcal{L}_{\text{obs}}(\delta, \psi_0, \psi_1) &= \prod_{i \in \mathcal{J}(1,0)} h_0(0; \mathbf{x}, \psi_{\mathbf{0}})(1 - g(\mathbf{x}, \delta)) \times \prod_{i \in \mathcal{J}(1,J+1)} g(\mathbf{x}, \delta) h_1(J; \mathbf{x}, \psi_{\mathbf{1}}) && (5) \\
&\times \prod_{y=1}^{J} \prod_{i \in \mathcal{J}(1,y)} \{g(\mathbf{x}, \delta) h_1(y-1; \mathbf{x}, \psi_{\mathbf{1}}) + (1 - g(\mathbf{x}, \delta)) h_0(y; \mathbf{x}, \psi_{\mathbf{0}})\} && (6) \\
&\times \prod_{y=0}^{J} \prod_{i \in \mathcal{J}(0,y)} \{g(\mathbf{x}, \delta) h_1(y; \mathbf{x}, \psi_{\mathbf{1}}) + (1 - g(\mathbf{x}, \delta)) h_0(y; \mathbf{x}, \psi_{\mathbf{0}})\} && (7)
\end{aligned}$$

A variety of distributional assumptions for $g(\cdot, \cdot)$ and $h_z(\cdot, \cdot)$ are possible. In the Monte Carlo exercises below I use the double binomial specification:

$$\begin{aligned}
g(\mathbf{x}, \delta) &= \text{logit}^{-1}(\mathbf{x}'\delta) && (8) \\
h_z(\mathbf{x}, \psi_z) &= J \times \text{logit}^{-1}(\mathbf{x}'\psi_z) && (9)
\end{aligned}$$

In the Monte Carlos here we also constrain $h_0(\mathbf{x}, \psi_0) = h_1(\mathbf{x}, \psi_1)$; this decision is inconsequential.

The likelihood described above is quite difficult to evaluate. The computational strategy

pursued in Blair and Imai (2010, 2012); Imai (2011); Imai, Park and Greene (2014) involves treating the $Z_i(1)$ as partially missing data and then deriving a much simpler complete data likelihood that can be maximized via the EM algorithm. This computational strategy is feasible *only* if the sets $\mathcal{J}(1,0)$ and $\mathcal{J}(1,J+1)$ are nonempty.

1.5 Random respondent error

The discussion in sections 1.2 and 1.3 makes clear the importance of the *no liars* assumption for our ability to extract more information from list experiment data. Line (5) instantiates that assumption in the likelihood in an important way: the sets $\mathcal{J}(1,0)$ and $\mathcal{J}(1,J+1)$ are those respondents in the treatment group who reveal with (assumed) certainty that, respectively, none and all of items on the treatment list apply to them. The composition of these sets not only affects these key terms in the likelihood but it also affects the composition of the sets $\mathcal{J}(1,y)$ in line 6. If these sets are empty then the ICT-MLE cannot be estimated.

Scholars using list experiments are sensitive to the assumption about truthfulness in respondents at the extremes, as our discussion of ceiling and floor effects shows. Ceiling and floor effects can be viewed as one type of (asymmetric) measurement error in which some respondents who should appear in the extreme categories chose not to report those true values. Ceiling effects imply a downward bias in $\hat{\pi}_{Z^*}$ regardless of the estimator. With this in mind, Kuklinski, Cobb and Gilens (1997) and Glynn (2013) provide survey design advice aimed at minimizing the subjects put in these positions. From the perspective of the ICT-MLE, ceiling and floor effects are situations of intentional respondent misrepresentation, situations that can be modeled. Blair and Imai (2012) provide a generalized ICT likelihood that incorporates a model for ceiling and floor effects.

Without detracting from the important work on ceiling/floor effects, it is worth highlighting that virtually[5] all work aimed at testing and relaxing ICT assumptions has focused

[5]Blair and Imai (2012) develop a simple test to identify departures from the *no design effects* assumption,

on strategic behavior by respondents, ignoring the implications of simple respondent error due to the usual problems of misunderstanding, rushing through surveys, and miscoding. The presence or absence of ceiling/floor effects tells us nothing about whether random error is also a concern, as we demonstrate in section 2. More importantly, the existing fixes for ceiling and floor effects treat the error as entirely asymmetric: we are willing to believe that some respondents strategically choose not to reveal an extreme value yet we also maintain the *assumption* that all the observed responses in the extreme categories are in fact truthful. The asymmetry and strength of these assumptions seem hard to justify.

The ICT-MLE relies heavily on the assumption that we can treat observed survey responses in the extremes of the treatment group distribution as truthful. But this assumption flies in the face of the concerns about respondent truthfulness that motivate the use of indirect questioning in the first place. Moreover, the parts of the response distribution needed to identify and estimate the model are almost certainly prone to very small sample sizes. In fact current survey design best practice involves taking steps to actively *minimize* the number of respondents that appear in exactly the cells required to identify and estimate the ICT-MLE. Both design objectives and the applied context for list experiments work against the assumptions underlying the ICT-MLE.

1.6 possible patterns of error and consequences

It is uncontroversial to assert that respondent/measurement error is endemic in surveys. The real questions surround the type of error and consequences for various estimators. In considering this it helps to pose some simple models of the error process. For example, one type of error, which I call *uniform error* would involve a process by 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 the respondent's T_i. Put another way, uniform random

which is not directly at issue here.

error will be correlated with the treatment indicator for the same reason that we expect heteroskedasticity in the DiM estimator: respondents in the treatment group have one more value $(J+1)$ in which to erroneously respond. We should therefore expect that uniform error induces bias and inconsistency in the DiM estimator resulting in an overestimate of π_{Z^*}. The degree of bias will depend, obviously, on the rate of error. Perhaps less obviously the longer the list the lower the correlation between treatment indicator and uniform respondent error. As $J \to \infty$ the bias problem disappears at the cost of increasing variance in the estimator.[6] The ICT-MLE will also be biased and inconsistent under uniform error for the obvious reason that the distributional assumptions that underpin the likelihood are incorrect. Uniform bias will result in more values in higher categories, on average, under treatment so the ICT-MLE will also over-estimate π_{Z^*}. Uniform error should be relatively innocuous compared to other types of error described below, however, because only $\frac{2}{J+1}$ of the erroneous responses in the treatment group will be treated as true observed values of Z^* under the no liars assumptions.

This simple discussion also has two important implications. First, whatever bias is induced by error will be exacerbated as the population prevalence of the sensitive item decreases. Since the error is correlated with treatment a decline in the underlying prevalence of the sensitive item implies that the observed difference between treatment and control will be increasingly driven by measurement error. For the ICT-MLE, a lower underlying frequency of the sensitive attribute implies there there will be fewer truthful-responders in the $\mathcal{J}(1, J+1)$ set. In the limit this set is composed entirely of noise, highlighting the estimator's sensitivity to assumptions about truthful response.

Second, the greater efficiency of the ICT-MLE under its assumptions, especially if covariate information is brought to bear, will generate relatively tight standard errors around a biased estimate if measurement error is a problem. The DiM estimator may be biased under

[6]To see the inutition here, let $e_{i0} \sim U[0, J]$ and $e_{i1} \sim U[0, J+1]$ be the discrete, uniform measurement error for the baseline and treatment groups, respectively. We then get $Y_i = \alpha + \tau T_i + e_{i1} T_i + e_{i0}(1 - T_i) + \epsilon_i$. As $J \to \infty$ $\mathrm{E}[e_1 - e_0] \to 0$ which implies that $\mathrm{Cov}(T_i, e_i) \to 0$.

uniform error but it is also noisier. The ICT-MLE, on the other hand, will not generate inflated standard errors under its maintained assumptions raising the risks of Type-I error, especially when sample sizes are smaller and the sensitive item is rare in the population.

Other error structures are obviously possible. I briefly consider three additional: *bottom-biased*, *top-biased*, and *extreme-biased*. By bottom-biased error I envision an error process by which the respondent's truthful response is randomly replaced with "0." Top-biased error is similar only replacing the truthful response with the maximum value available to her. Extreme-biased error would combine the two.

Bottom-biased error is interesting in that it is weakly correlated with treatment. The DiM estimator should therefore remain nearly unbiased. The ICT-MLE, however will treats all "0" responses in the treatment group as revelations of the absence of the sensitive attribute and will therefore rely heavily on them in identification and estimation. Erroneously treating respondents in the $\mathcal{J}(1,0)$ group as truthful will downwardly bias the ICT-MLE $\hat{\pi}_{Z^*}$. We expect to see relatively little effect of bottom-biased error on DiM whereas there should be substantial problems with ICT-MLE.

Top-biased error is likely to be the most problematic for both the DiM and ICT-ML estimators. It is correlated with treatment, by construction, and this relationship will not weaken as the list length grows. Errors in $\mathcal{J}(1, J+1)$ will again present serious problems for the ICT-MLE as the observed "$J+1$" responses are all treated as truthful and the number of responses in this set should be small if we are following question design best practices. Top-biased error should lead to severe over-prediction of π_{Z^*} for both estimators but I conjecture that the ICT-MLE will perform worse, especially when the underlying rate for the sensitive attribute is low.

Extreme-biased error should split the difference between top- and bottom-biased error. It will induce bias in both estimators, but not as badly as purely top-biased error. It will be more problematic than uniform error.

To be clear these error structures are thought experiments; I am not asserting that any describe actual behavior. Indeed I anticipate that any real population would exhibit a mixture of error types. If we were to somehow correctly assume a particular error process and its rate then it might be modeled. But having such knowledge in an applied setting seems unlikely. Trying to model error by making more and stronger assumptions seems like a high-cost, low-return strategy for addressing the measurement problem at hand.

2 An example

To demonstrate that this problem of respondent error and estimator bias is more than theoretical, I rely on the list experiments reported in Ahlquist, Mayer and Jackman (2014), henceforth AMJ. AMJ use a YouGov Internet panel to ask questions about voter impersonation and vote buying in the 2012 US election. The sample was split such that the control group for the vote buying question served as the treatment group for the voter impersonation question and vice versa.

AMJ find no evidence of either impersonation or fraud. Nevertheless, there are some anomalies. First, the DiM estimates and ICT-MLE estimates are noticeably different, with the ICT-MLE showing point estimates substantially larger than those using the simple mean comparison procedure. Second, a non-negligible proportion of respondents (about 2.5% of the treatment sample, twelve individuals) in the voter impersonation question claimed the maximum number of items (5), thereby admitting to voter impersonation (in addition to a variety of other things) if we assume these responses are truthful, as the ICT-MLE does. One could therefore construe this 2.5% as a lower bound estimate for the rate of voter impersonation. If this estimate were true then the survey implies that *at least* five million people cast fraudulent ballots in the 2012 election—a shocking number inconsistent will all other work on this topic.

After examining the broader survey behavior of the respondents who claimed the maximum of five in the treatment condition for the voter impersonation question, AMJ find further reason to treat these responses as suspect. Most of those choosing the maximum value in the list experiments, whether in the treatment or control groups, appeared to be rushing to complete the survey as fast as possible, not revealing actual behaviors.

To further investigate this conjecture AMJ fielded a second set of list experiments in September 2013 with a new YouGov sample.[7] In addition to replicating the original list experiment questions they fielded two more list experiments as calibration exercises. The first new question offered subjects the opportunity to admit to something believed to occur with (near?)zero probability: abduction by extraterrestrials. The alien abduction list experiment is described in appendix table 4. The second of the new list experiments asks respondents about a common behavior that is illegal in most states: sending or reading text messages while driving. The details of this question are found in appendix table 5. AMJ found two previous large surveys on the subject of texting and driving. Madden and Rainie (2010) find that that 27% of US adults have sent or read a text message while driving. Naumann (2011) reports the results of cross-national surveys which estimates that about 31% of U.S. drivers aged 18-64 years had sent an SMS while driving in the last 30 days. Both surveys used direct questioning techniques.

Figure 2 displays population prevalence estimates for all four list experiments as calculated using both DiM and ICT regression.[8] Several things are immediately apparent. First,

[7] $N = 3000$, three times the size of the original sample. Using the test proposed by Blair and Imai (2012) there was no evidence leading to the rejection the null hypothesis of no design effect for any of these questions.

[8] I use the double-binomial maximum likelihood estimator here and ignore survey weights reported in the initial paper. In fitting the ICT regressions we included age, race, and gender as covariates. We also fit models exploring ceiling and floor effects. In all cases where the floor/ceiling models converged there was never any reason to prefer them over the simpler versions based on likelihood ratio tests. Finally, lest there be worry that some respondents are intentionally responding to the abduction question as a lark, note that AMJ pretested the alien abduction questions against an alternative which read "I won more than $1 million in the lottery." Findings were very similar between the two.

the list experiment examining a relatively common behavior recovers rates of texting-while-driving in line with previous estimates. The ICT and DiM estimates are very close to one another and the uncertainty around the ICT estimates is substantially narrower, reflecting the efficiency improvement in the ICT-MLE, bought with distributional assumptions and the incorporation of covariate information. But when we turn to the low-prevalence questions (impersonation and abduction) there are massive differences between the ICT-MLE and DiM estimates. The DiM estimates for both voter impersonation and alien abduction are very close to zero, consistent with both prior expectations and the earlier survey wave. The ICT estimates are shockingly large and have relatively narrow standard errors, raising the prospect of erroneous inference. Clearly something is happening to degrade the performance of the ICT-MLE when the underlying prevalence of the sensitive item is low.

AMJ then go on to look at proportion of respondents in the treatment groups claiming the maximum possible number of items (i.e., the sets $\mathcal{J}(1,5)$). Table 2 displays their findings. The proportion of people answering the maximum is remarkably stable, around 2-3%, even for sensitive behaviors that are far more common in the population (texting while driving). The rate of 2.4% is especially remarkable for the alien abduction question, since answering "5" in that context corresponds to admitting to alien abduction *and* serving on jury duty *and* IRS auditing. The rate of IRS auditing in FY2013 was 0.96% (Internal Revenue Service, 2014), which implies that *at least* 60% of the respondents in $\mathcal{J}(1,5)$ for the alien abduction question are likely erroneous responses. Moreover, of those answering "5" for alien abduction, 24% (9/37) also answered "5" for voter impersonation.

All this leads to two conclusions. First, there is non-negligible respondent error in these data, as we would expect with any real-world survey. Second, the ICT-MLE vastly overestimates the prevalence of two sensitive attributes, both of which have very low (0?) population prevalence.

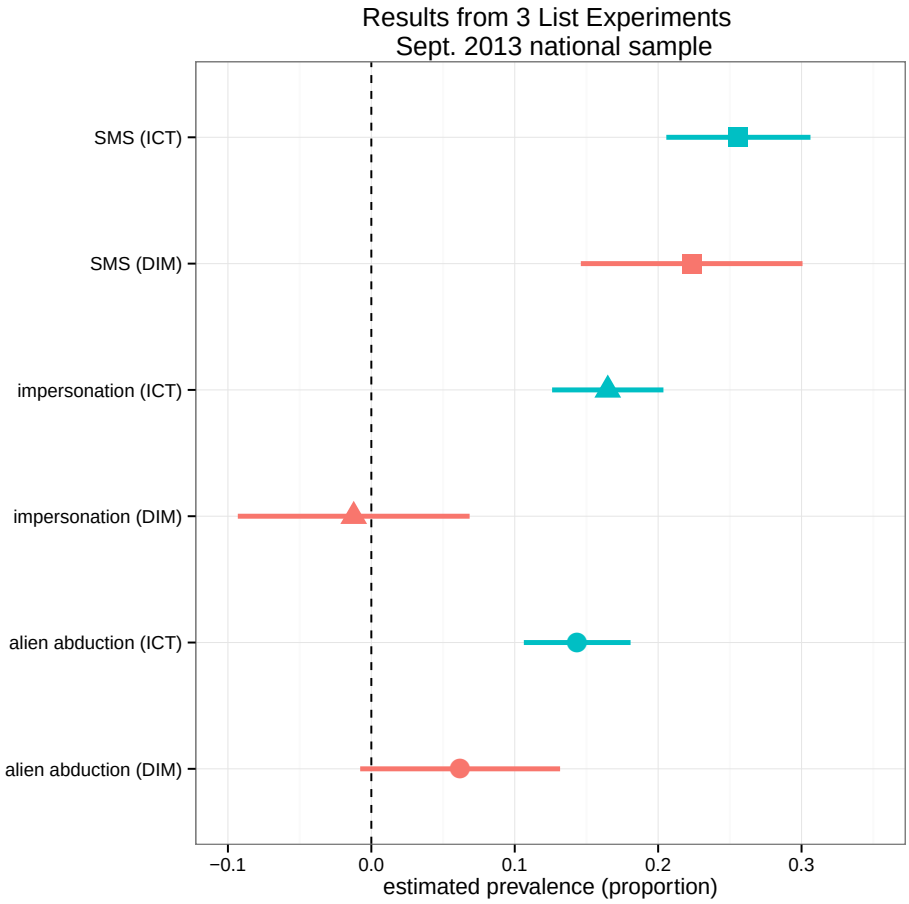

Figure 2: Evidence of problems with the ICT regression estimator. We ignore survey weights here. Bars are 95% CIs for the DIM estimates and ±2SEs for the ICT estimates. DIM = difference-in-means, ICT= Item Count Technique regression.

3 Monte Carlo Experiments

To gain better purchase on sensitivity of the DiM and ICT-ML estimator to respondent error I conduct a series of Monte Carlo experiments in which I vary the underlying prevalence of the sensitive item and the rate and type of respondent error. I evaluate the simulations on the following dimensions:

- Computational stability

Table 2: The proportion of respondents selecting the most extreme value is very stable across survey waves and questions. Source: Ahlquist, Mayer and Jackman (2014).

	Wave	% treated choosing "5"	treated N
Voter impersonation	Dec. 2012	2.5%	486
Voter impersonation	Sept. 2013	2.7%	1528
Alien abduction	Sept. 2013	2.4%	1528
texting while driving	Sept. 2013	3.3%	1472

- Bias in the estimate of the population prevalence of the sensitive item
- Coverage rates of estimated 95% confidence intervals for population prevalence estimates
- Bias in covariate parameter estimates

3.1 Monte Carlo design and notation

We begin with a fixed population, defined by eight attributes. Four binary attributes denoted $\{C_1, C_2, C_3, C_4\}$, and four regression parameters, b_1 and $\{b_0^L, b_0^M, b_0^H\}$. The first four are used to generate a hypothetical respondent's values for each of four "control" items whereas the last four are used to determine a respondent's value on each of three "sensitive" items occurring with low, medium, and high rates, respectively, in the population. In other words, we use these population parameters to generate the latent values that underly the responses to three hypothetical list experiments, each having the same $J = 4$ control items.

Each individual's probability of answering affirmatively for each of the three sensitive items is given by

$$\begin{aligned} \Pr(Z_i^k = 1) &= \text{logit}^{-1}(b_0^k + b_1 x_i), \quad k \in \{L, M, H\} \\ X &\sim N(0, 1) \end{aligned} \tag{10}$$

The only covariate is X. I fix $b_1 = 1.5$ throughout. I choose values of b_0^k such that the underlying population prevalences are $\pi_{Z^*}^L = 0.02$, $\pi_{Z^*}^M = 0.10$, and $\pi_{Z^*}^H = 0.25$.

The population-level parameter values for each of the attributes are displayed in table 3. Note the structure of the control items. C_3 and C_4 are relatively low- and high-frequency attributes, respectively, designed to avoid ceiling and floor effects. C_1 and C_2 are common, each appearing in half the population, but with a moderate negative correlation, following the discussion in Glynn (2013). In other words, the underlying frequencies of the control items are designed to make the list experiment as fruitful as possible.

Table 3: Key population-level parameters for the Monte Carlo experiments

Attribute	True rate	Comments
C_1	0.50	$\text{cor}(C_1, C_2) = -0.6$
C_2	0.50	$\text{cor}(C_1, C_2) = -0.6$
C_3	0.15	independent of other attributes
C_4	0.85	independent of other attributes
b_1	1.5	covariate parameter
b_0^L	-4.92	intercept for $\pi_z = 0.02$
b_0^M	-2.96	intercept for $\pi_z = 0.10$
b_0^H	-1.54	intercept for $\pi_z = 0.25$
μ	0	population mean for covariate X
σ	1	population standard deviation for covariate X

For each of the 1000 Monte Carlo runs I generate a sample of 1000 "respondents", generating actual values for $C_1, \ldots, C_4$ and z_i^k based on the parameters just described. With equal probability we randomly assign each of the respondents, i, to be in the treatment group or the control group, denoted by binary variable T_i. For each of the respondents we

then calculate the the error-free observed outcome for $k \in \{L, M, H\}$ as

$$\begin{aligned} Y_i^k \mid T^i = 0 &= \sum_{j=1}^{4} C_{ij} \\ Y_i^k \mid T^i = 1 &= z_i^k + \sum_{j=1}^{4} C_{ij} \end{aligned}$$

That is, Y^k represents the data we would hope to observe in a well-designed list experiment with no measurement error and with respondents satisfying Imai's three basic identification assumptions.

We then introduce four different kinds of respondent error, each of which occurs at two different rates. The two rates of error, r, are 0.02 and 0.06. The 2% rate is approximately what was observed by Ahlquist, Mayer and Jackman (2014) whereas 6% represents a high respondent error rate. The four kinds of respondent error induced here are:

- **Uniform error**: We randomly select $1000r$ respondents. For each of the selected respondents, if $T_i = 0$ we replace Y_i^k with a random draw from $\{0, 1, 2, 3, 4\}$ for each k. If $T_i = 1$ we do the same thing only drawing from $\{0, 1, 2, 3, 4, 5\}$.[9]

- **Top-biased error**: We randomly select $1000r$ respondents. For each of the selected respondents, if $T_i = 0$ we replace Y_i^k with a 4 for each k. If $T_i = 1$ we do the same thing only with a 5.

- **Bottom-biased error**: We randomly select $1000r$ respondents. For each of the selected respondents we replace Y_i^k with a 0.

- **Extreme-biased error**: We randomly select $1000r$ respondents. For each of the selected respondents, if $T_i = 0$ we replace Y_i^k with a random draw from $\{0, 4\}$ for each k. If $T_i = 1$ we do the same thing only drawing from $\{0, 5\}$.

[9]Note that for a selected respondent we insert the same replacement value for each of the three Y_i^k.

Note that while all these forms of measurement error violate Imai's *no liars* assumption, they still leave us with data that satisfy the *randomization* and *no design effects* assumptions.[10] Thus we can view this study as examining the estimators' sensitivity to violations of the no liars condition.

For each of the error types and rates within each Monte Carlo iteration we fit two models. The first is the ICT-MLE with the double-binomial structure described above. I include X as the only covariate. The second is the DiM, calculated as the OLS regression

$$Y_i^k = \alpha^k + \tau^k T_i + \beta^k x_i + \delta^k T_i x_i \tag{11}$$

Following Imai (2011) we calculate heteroskedasticity-corrected standard errors for the OLS regression.[11]

3.2 Results

The Monte Carlo results presented here took over 18 days to run in $\mathcal{R}$ 3.1.0 on remote Linux-based server cluster using `list` v. 7.0.

3.2.1 computational stability

We first consider the computational stability of the ICT-MLE.[12] Recall that the ICT-MLE implemented in Blair and Imai (2010) relies on the EM algorithm to maximize an observed-data likelihood, treating responses to the sensitive item as partially missing. The observed Z_i are derived from the $\mathcal{J}(1,0)$ and $\mathcal{J}(1,J+1)$ responses.

EM is known to be slow to converge or unstable when the amount of missingness is

[10] Were we to restrict measurement error to only occur among the $T_i = 0$ group we could maintain the no liars assumption at the cost of violating the no design effects assumption.

[11] We also fit models with only an intercept and treatment indicator for the purposes of the Hausman test discusses below.

[12] The DiM estimator had no computational difficulties.

extreme relative to the observed values. The lower the prevalence of the sensitive item the correspondingly fewer observed cases of $Z_i = 1$ and the less stable we expect the algorithm to be. Moreover, inducing error into the system will inflate the number of cases in $\mathcal{J}(1, J+1)$; this will obviously be most pronounced under top- and extreme-biased error. We therefore expect the ICT-MLE to be most unstable in the low prevalence, no error condition. The instability will be mitigated most rapidly by inducing error that skews into the top of the response distribution.

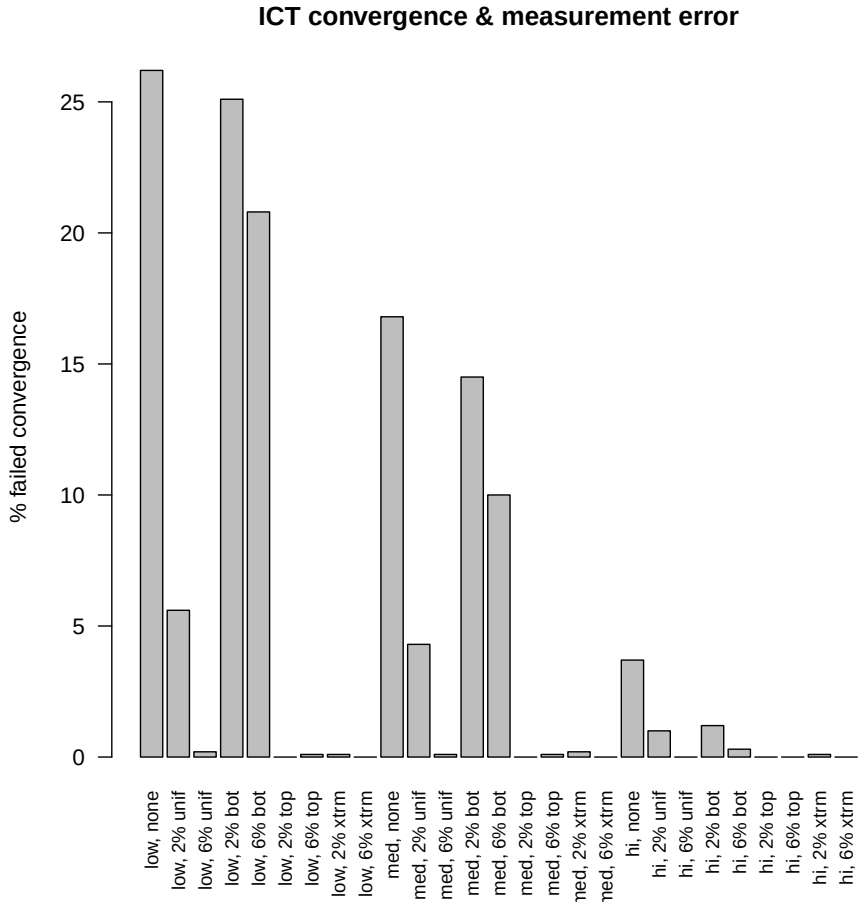

Figure 3: Percent of Monte Carlo runs in which the ICT-ML estimator was unable to run appropriately. A low prevalence sensitive item makes the ICT-MLE less computationally stable. More error, especially when biased into higher categories enables EM to run more easily.

Figure 3 displays the results from the Monte Carlo experiments. The ICT-MLE becomes increasingly fragile as the underlying prevalence of the sensitive item decreases. In the low-prevalence (2%) condition the algorithm failed to converge or otherwise kicked out an error more than 25% of the time. This error rate declined markedly in the medium-prevalence condition and vanished in the high prevalence condition. As shown in figure 3, inducing error has a similar effect on the stability of the algorithm; error at the top extreme, even at low levels, prevented convergence problems. This has some problematic implications: when the underlying behavior of interest is rare but respondents are answering truthfully the algorithm is less stable, but when there is error the estimator is more likely to return an answer, but one that is biased (as we shall see).

3.2.2 bias in population prevalence estimates

Figure 4 presents the baseline findings with no measurement errors induced. The solid line represents the distribution of the ICT-MLE error across the 1000 Monte Carlo simulations while the broken line is the simple difference-in-means estimator.

In the absence of error, the difference-in-means estimator is unbiased. Unexpectedly and contrary to the simulation studies by Imai (2011) and Blair and Imai (2012), the ICT-ML estimator is showing bias even in the no-error setting. In both the low- and medium-prevalence condition the ICT-MLE is more efficient, but around the wrong value. While it may be argued that the sample size of 1000 is insufficient for the asymptotics of the MLE to kick in we note that most mass surveys are conducted with samples around 1000 (see, e.g., Rosenfeld, Imai and Shapiro (2015)), so $N = 1000$ is a relevant threshold. Subsequent trials with $N = 2000$ yielded similar results.

Figure 5 displays the distributions of bias for the uniform error situations. At low error levels the DiM estimator is nearly unbiased regardless of the underlying prevalence of the sensitive item. At higher error rates the DiM estimator is overestimating the true prevalence,

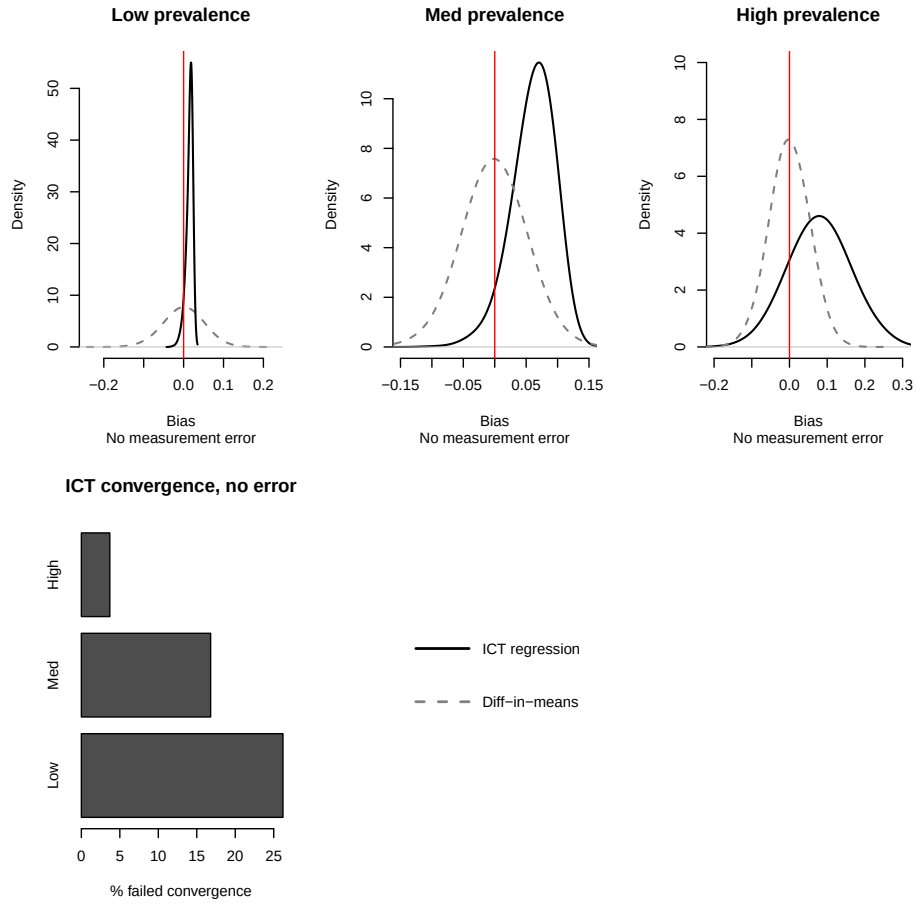

Figure 4: Bias in the ICT-ML (solid) and difference-in-means (broken line) estimators with no respondent error. Curves display the distribution of estimates across 1000 Monte Carlo simulations. When the sensitive item is rare the algorithm has trouble converging; error, especially at the top end, enables the algorithm to converge–but at the wrong answer.

as expected. The size of bias diminishes as the sensitive item becomes more common.

The ICT-MLE also seems affected by uniform error, even at low error levels, but not in a consistent way. Uniform error appears to interact with the underlying prevalence to shift the ICT-ML estimator in different directions. Again the ICT-MLE is showing less variance than the DiM estimator.

Figure 6 presents a key finding: top-biased error severely upwardly biases the ICT-ML estimator, even at low error levels. This bias is massive when the underlying prevalence of

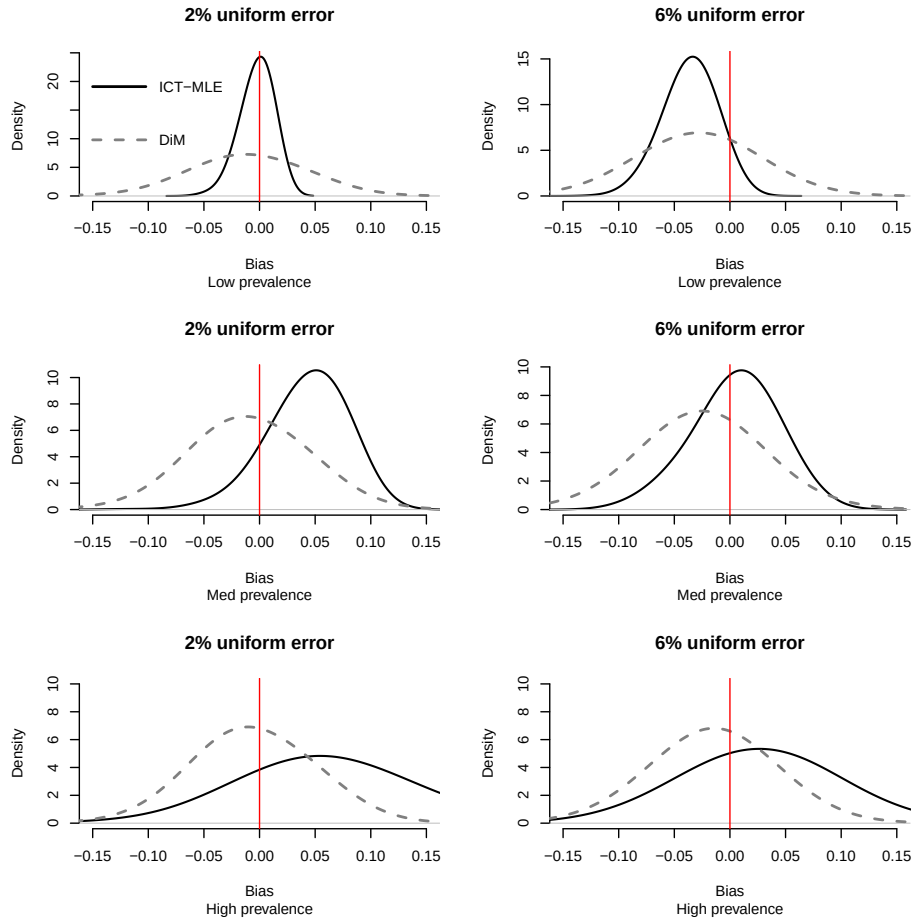

Figure 5: Bias in the ICT-ML (solid) and difference-in-means (broken line) estimators under random respondent error across differing underlying frequencies of the sensitive item. Curves display the distribution of estimates across 1000 Monte Carlo simulations.

the sensitive item is low: under $r = 6\%$ the ICT model puts the prevalence of the sensitive item at a level $10\times$ greater, on average, than the true value (21% as opposed to 2%). The DiM estimator also shows upward bias, as expected, but not nearly to the same extent as the ICT-MLE. As the underlying prevalence increases the severity of the bias in both estimators diminishes, but the ICT-ML estimator remains the more affected.

Figures 7 and 8 look at bottom- and extreme-biased error, respectively. As expected the DiM estimator remains unbiased under bottom-biased error. The ICT-MLE, however,

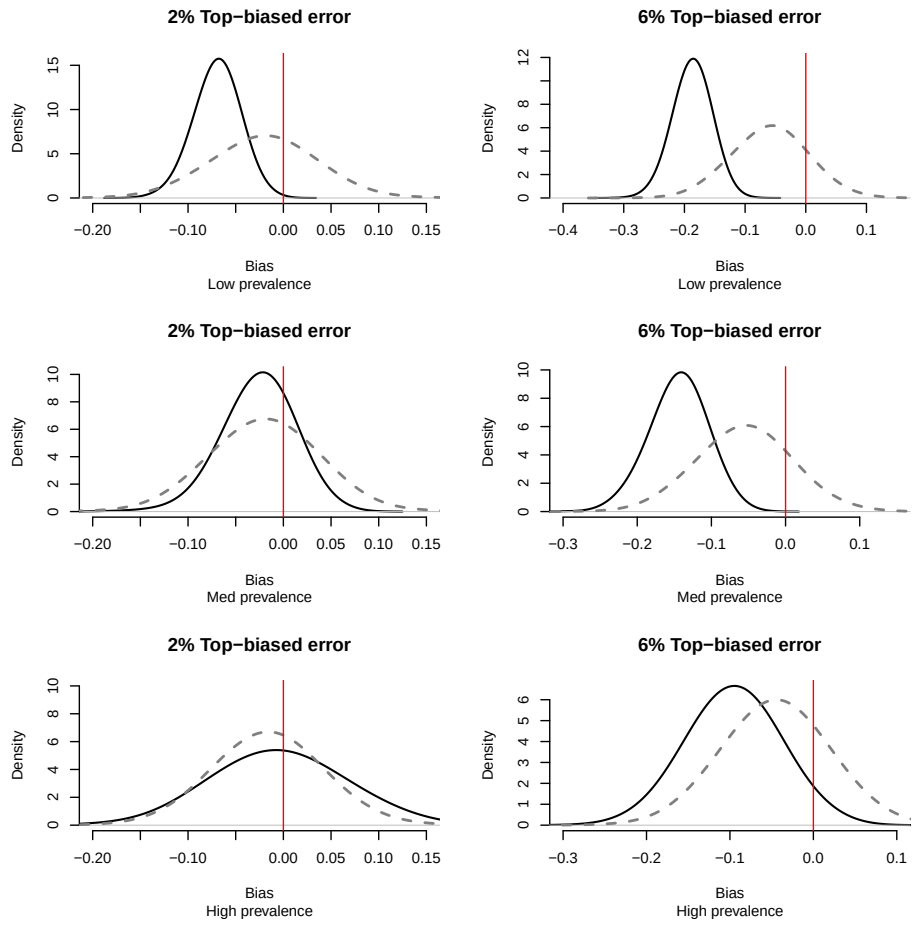

Figure 6: Bias in the ICT-ML (solid) and difference-in-means (broken line) estimators under top-biased respondent error across differing underlying frequencies of the sensitive item. Curves display the distribution of estimates across 1000 Monte Carlo simulations.

exhibits strong downward bias. Bottom-biased error causing the ICT-MLE to underestimate the population prevalence of the sensitive item by 5-10 percentage points under medium prevalence and 10-20 percentage points under high prevalence. Extreme-biased error, as shown in figure 8, continues to induce bias in the ICT-MLE, but, as one would expect, the bias does not reach the same levels as with pure top- or bottom-biased respondent errors.

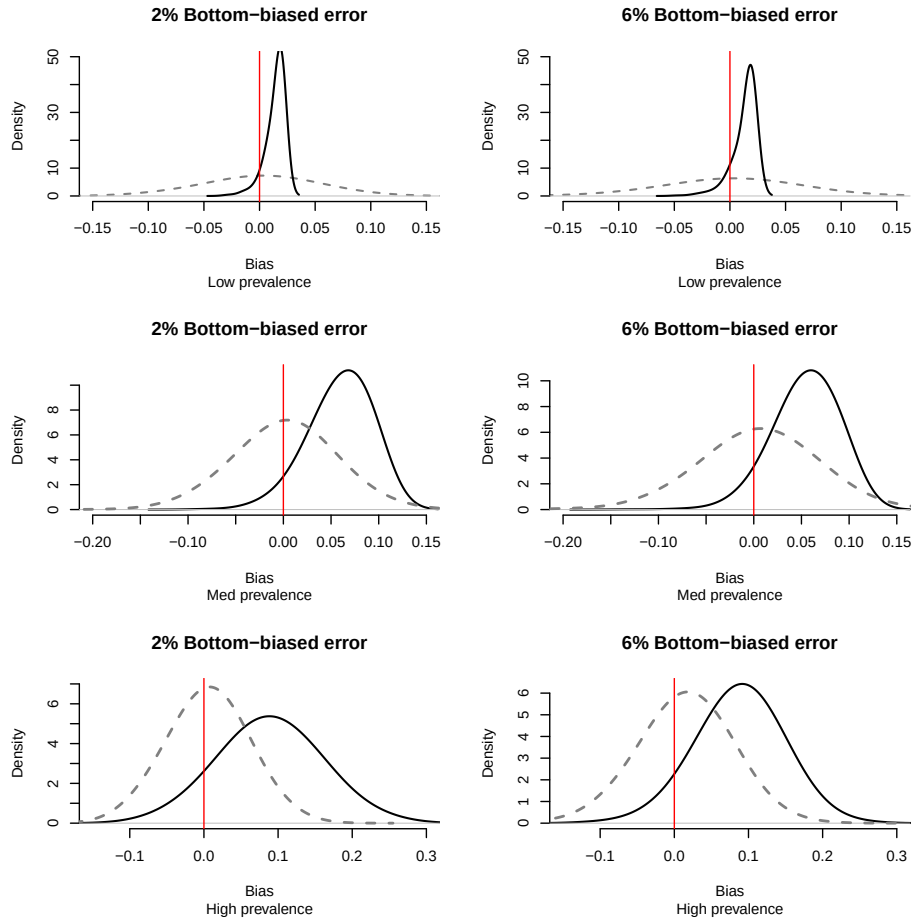

Figure 7: Bias in the ICT-ML (solid) and difference-in-means (broken line) estimators under bottom-biased respondent error across differing underlying frequencies of the sensitive item. Curves display the distribution of estimates across 1000 Monte Carlo simulations.

3.3 Standard error estimates & CI coverage rates

The estimated standard errors around the $\hat{\pi}^k_{Z^*}$ become increasingly important given the bias in the ICT-MLE just discovered. Small standard errors combined with biased estimates raise the risk type-I errors. To evaluate the performance of the two estimators, I construct two sets of displays. In figure 9 I compare the standard deviations of the estimated $\hat{\pi}^k_{Z^*}$ across the 1000 Monte Carlo runs against the standard errors for $\hat{\pi}^k_{Z^*}$ calculated at each Monte Carlo iteration, averaged across the 1000 iterations. That is, the figure reports $\text{SD}(\hat{\pi}^k_{Z^*}) - \overline{\widehat{\text{SE}}(\hat{\pi}^k_{Z^*})}$

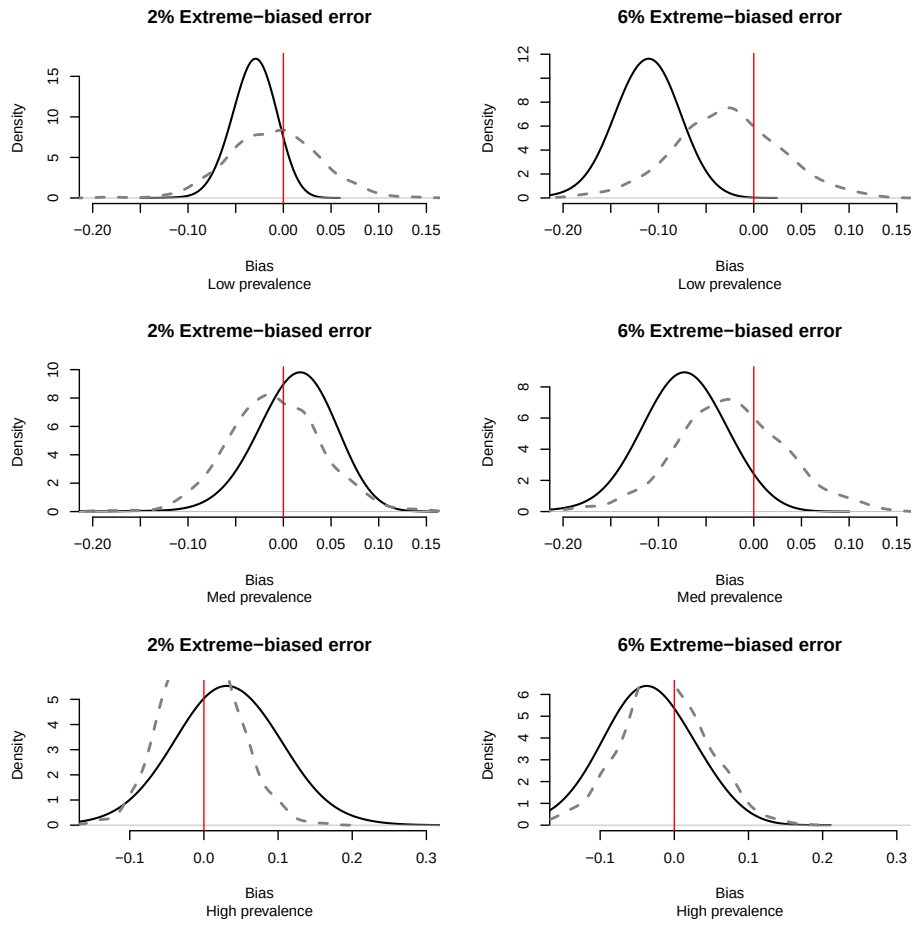

Figure 8: Bias in the ICT-ML (solid) and difference-in-means (broken line) estimators under extreme-biased respondent error across differing underlying frequencies of the sensitive item. Curves display the distribution of estimates across 1000 Monte Carlo simulations.

as a way of describing the bias in the ICT-ML and DiM estimates of their respective sampling distributions.

The results are stark. The DiM standard error estimates are unrelated to the underlying prevalence of the sensitive item in all conditions. They are unbiased in the no error and bottom error conditions. Under uniform error DiM does begin to overestimate the standard errors; this gets worse under extreme error and worst under top-biased error. But the problems in the DiM estimates are minor compared to the ICT-MLE. The ICT-MLE bias in

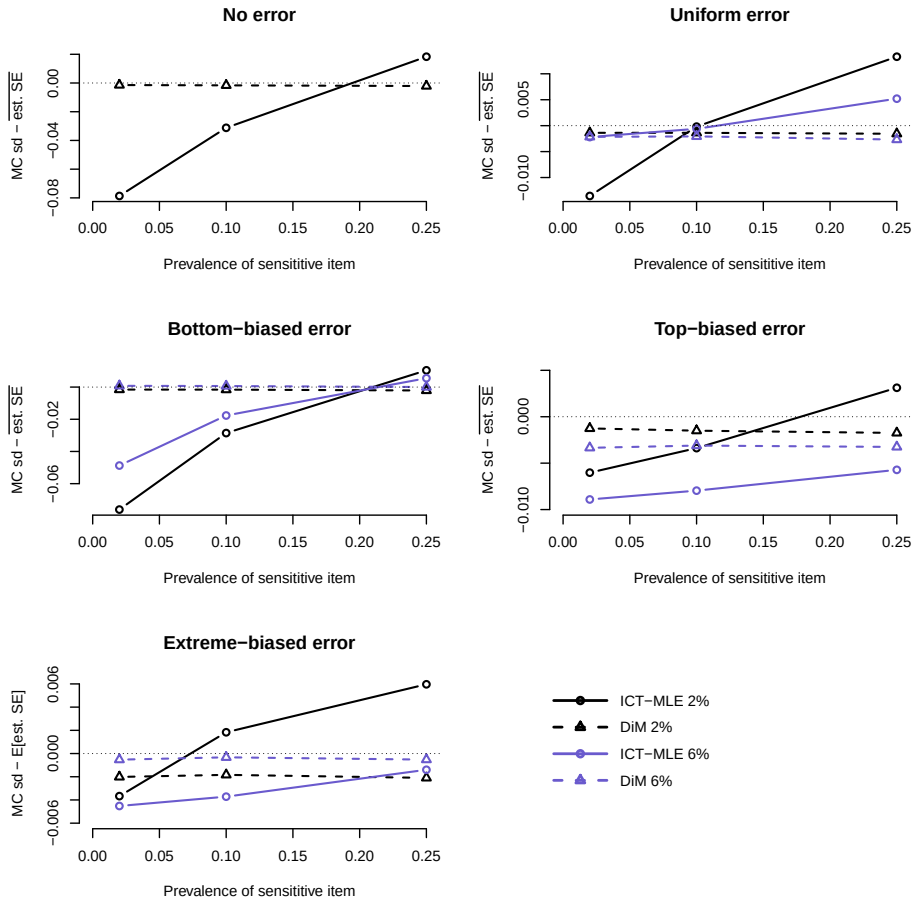

Figure 9: Difference between the standard deviation of Monte Carlo population prevalence estimates minus the mean estimated standard error for prevalence estimates across 100 Monte Carlo runs ($\mathrm{SD}(\hat{\pi}^k_{Z^*}) - \overline{\widehat{\mathrm{SE}}(\hat{\pi}^k_{Z^*})}$). DiM standard errors are heteroskedasticity-corrected.

standard error estimates are significant, even under the no error condition and show a strong positive correlation with π_{Z^*} in all conditions. When the population prevalence is high the ICT-MLE is underestimating standard errors and when it is low it is overestimating them. Error, especially bottom and top biased exacerbate this relationship.

In the second set of displays, figure 10, I calculate coverage rate: the percent of simulations in which the 95% nominal confidence interval for $\hat{\pi}^k_{Z^*}$ actually contains $\pi^k_{Z^*}$. Again the results are stark. The DiM estimator performs as expected: coverage is unrelated to the underlying

prevalence of the sensitive item for all conditions. Under the no error and bottom-biased conditions the confidence intervals perform as they should. Top and extreme biased error give 95% CIs that are somewhat too narrow. Again, however, the problems with the DiM estimator are minor in comparison to the ICT-MLE. Even under no error the standard errors are far too narrow. Top- and extreme-biased errors produce exactly what we feared: estimates are badly biased and standard errors are too small such that virtually *none* of the ICT-MLE estimated confidence intervals contained the true value when the underlying prevalence is low.

3.4 Covariate parameter estimates

The ICT-MLE was specifically built to include covariates. We therefore included the covariate, X, in the Monte Carlo experiments. Using the double-binomial ICT-MLE specification we can compare the estimated coefficient from the $g(\delta x)$ part of the ICT-ML model in equation 8 to $b_1 = 1.5$ from equation 10. For the DiM estimator we calculate bias in the covariate parameter by differencing the marginal effect of X on $\Pr(Z_i = 1 \mid X_i = 0)$ implied by equation 10 from the $\hat{\delta}$ defined in equation 11.[13]

Results are displayed in figure 11. Again we see bias in the ICT-ML estimate, this time for covariate parameters. Even in the no error setting ICT-MLE is systematically overestimating the covariate parameter by anywhere from 5-30%. Note that the accuracy of the ICT-MLE is strongly correlated with the underlying prevalence of the sensitive item. This relationship can be viewed as a rare events logit problem (King and Zeng, 2001): there are very few observed "1s" in the low prevalence context, inducing bias that rapidly decreases as the event becomes more common. Interestingly, error (except the bottom-biased variety) seems to actually improve the performance of the ICT-MLE, reducing bias in the low-prevalence situations. This is an artifact of the simulation design: error is uncorrelated with X by

[13]The implied marginal effect evaluated at $\mathrm{E}[X] = 0$ is given by $b_1 \exp(b_0^k)/(1 + \exp(b_0^k))^2$.

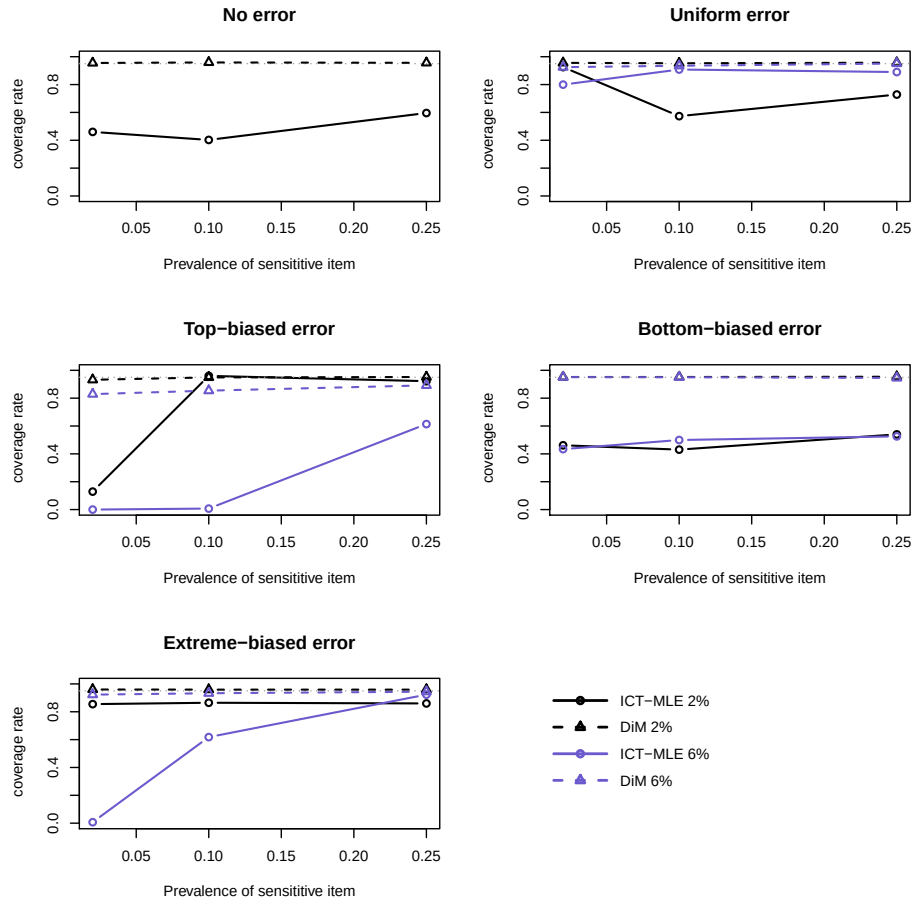

Figure 10: Coverage rates for 95% nominal confidence intervals for π_z based an assumed Normal sampling distribution. ICT-MLE (solid) and DiM (broken line) estimators under differing underlying frequencies of the sensitive item and differing error rates. DiM standard errors are heteroskedasticity-corrected.

construction and it artificially inflates the number of observed "1s", thereby mitigating the rare events problem. DiM estimates are slightly sensitive to π_{Z^*} but in general show much less bias than the ICT-MLE. Even in the domain that motivates the ICT-MLE—the inclusion of covariates—we see that the estimator is returning biased answers in some cases and showing much more sensitivity to respondent error than the DiM/OLS approach.

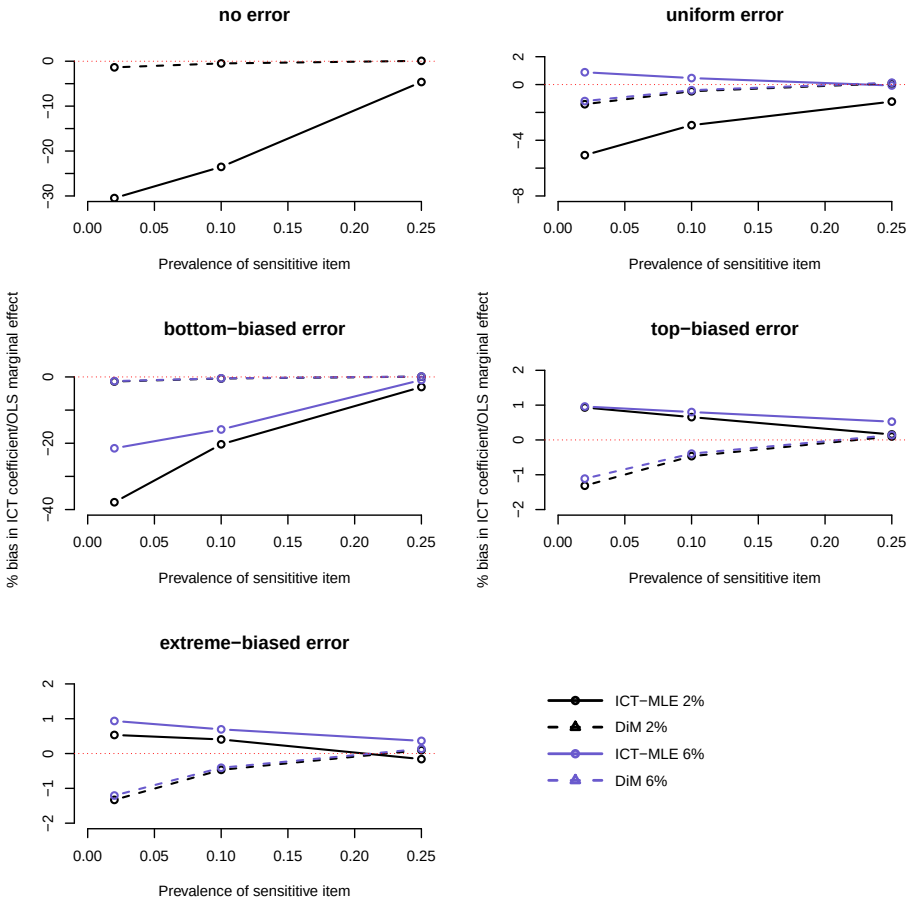

Figure 11: Bias in covariate parameter estimates as % of true parameter value for different estimators, error types and error rates. For DiM estimates we compare the regression parameter estimate to the marginal effect of X on $\Pr(Z_i = 1 \mid X_i = 0)$ implied by equation 10.

4 Implications and working toward solutions

While error can cause problems for both the DiM and ICT-MLE, the Monte Carlo simulations show the ICT-MLE's sensitivity to violations of the no liars assumption. The DIM estimator is the more robust in all cases. Problems with bias are compounded as the sensitive attribute becomes rarer in the population.

These findings lead us to consider—and reject—some possible solutions to the problem and then to consider when biased error is more likely to arise. We conclude with some advice

to applied researchers considering list experiments.

4.1 Solutions that will *not* work

4.1.1 Hausman Test

If the ICT distributional assumptions are correct then both the difference-in-means and the ICT estimators are consistent, but the ICT estimator, as a maximum likelihood estimator, is the more efficient. If the ICT distributional assumptions are not met then ICT estimator is no longer consistent while the difference-in-means estimator is. This represents a special (scalar) case of the Wu-Hausman test.

$$H = (\hat{\pi}_{ICT} - \hat{\pi}_{DiM})(\widehat{\text{Var}}(\hat{\pi}_{DiM}) - \widehat{\text{Var}}(\hat{\pi}_{ICT}))^{-1}(\hat{\pi}_{ICT} - \hat{\pi}_{DiM}) \quad (12)$$

$$H|_{h_0} \sim \chi^2_1 \quad (13)$$

where $\hat{\pi}_{ICT}$ and $\hat{\pi}_{DiM}$ are the estimated population prevalences of the sensitive item from the ICT and difference-in-means estimators, respectively, and $\widehat{\text{Var}}(\hat{\pi}_{DiM}), \widehat{\text{Var}}(\hat{\pi}_{ICT})$ are the estimated variances.

Unfortunately, the Monte Carlo results show that the ICT-MLE is not generating unbiased estimates of either the quantity of interest or the variance around that estimate. As a result a Hausman specification test will not be yield useful findings.[14]

4.1.2 CACE analysis/instrumental variables

List experiments are measurement devices, not actual experiments designed to identify a causal effect. But we might think to view list experiments through the lens of treatment noncompliance. Being in the treatment group is now an attempt to "treat" a respondent

[14]We also estimated the Hausman test statistic in the Monte Carlos just described. Monte Carlo results confirmed the Hausman test performed poorly. Results are omitted here for space considerations.

who may or may not comply (answer truthfully) with the treatment regimen. Analyzing experiments under noncompliance typically requires that we have information on those in the treatment group who actually comply with treatment and assume monotonicity in treatment assignment in order to estimate the complier average causal effect (CACE) using instrumental variables (Angrist, Imbens and Rubin, 1996). In the context of the list experiment this information on compliance among the treated is obviously missing. Moreover, even if this information were available we are not actually interested in the CACE. We are interested in the population prevalence of the sensitive item, the list experiment estimate of which is composed of those who will truthfully reveal their status in the treatment arm on the survey ("compliers") as well as those who do not. This latter group is composed of both random error and those who will never reveal their status. Successfully estimating a CACE, even if this were possible, would, at best, yield a lower bound on π_{Z^*}.

4.2 Situations likely prone to error

This sensitivity of the ICT-MLE to different error types and the less pronounced but still worrisome bias in the DiM estimator begs the questions of when different sorts of respondent error are more likely to arise. An obvious variable to consider is survey mode. I conjecture that there are likely to be risks associated with Web-administered survey that are particularly pernicious under the ICT-MLE framework. I hasten to add that my conjectures are, at this point, untested and warrant future research.

Incentivized, Internet-based panels like YouGov would seem to reduce some risks of respondent error while possibly amplifying others. On the positive side of the ledger, Internet-based surveys provide a level of anonymity and noninvasiveness that in-person or phone-based surveys cannot readily match (Gooch and Vavreck, forthcoming). The greater perceived anonymity of Web-based or other self-completed surveys may reduce the severity of list experiment design and floor/ceiling effects. On the negative side, however, Internet-based

surveys seem particularly likely to generate respondents rushing through the survey for at least two reasons. First, there is no interviewer who can control the pace of survey administration (though one can imagine several technical interventions to mitigate this problem). As a result we may get more noise, in general, with an Internet-administered survey. Second, at least in the YouGov example just described, it is easy for respondents to simply leave their cursors in one place and rapidly click through the survey; the AMJ analysis seems to indicate that this is more than just a passing concern. Depending on the positioning of click boxes on the screen and the defaults of both the survey and the user's Internet browser, such behavior could potentially lead to a variety of possible types of respondent error, including error systematically biased toward the extreme answers. Surveys taken on touchscreen devices will almost surely be even more prone to such problems. Combining these mode-based challenges with the ICT model's sensitivity to respondent error at the extremes could make Internet panels a dangerous environment in which to deploy such analysis tools. Systematic investigation of survey mode effects for indirect questioning has yet to occur.

4.3 Some advice

Unfortunately we do not yet have tools for determining the levels, rates, and structure of respondent error. But we can offer some advice.

List experiments are poor tools for finding rare events and behaviors. Contrary to conclusions in Kiewiet de Jonge and Nickerson (2013), which rely on real surveys and not controlled Monte Carlo studies, survey list experiments are likely to be poor tools for reliably estimating small values of π_{Z^*}. This is not surprising: mass surveys are notoriously weak at establishing the prevalence of rare attributes even when direct questioning is reasonable. List experiments are even less effective in that regard, but the likely bias in the list experiment analysis tools, especially ICT-MLE, provides an additional reason for caution. The DiM estimate (which preformed adequately in the Kiewiet de Jonge and Nickerson (2013) studies) should be the

initial point of departure.

Some would argue that list experiments (or even mass surveys) should be avoided entirely if we have prior beliefs that the sensitive attribute is rare. This seems a step too far. If we already knew knew the true population prevalence we would not need to run a survey. If the attribute of interest were easy to talk about we could be more confident in our priors and would not feel the need to employ indirect questioning. Realistically, as in AMJ, there will be attributes of interest that are arguably worth investigating with list experiments that turn out to be very rare in the population. Applied researchers need to recognize that available tools are fragile in such situations.

Adjust survey administration to minimize error. This advice seems trite, so let me offer more specific suggestions. First, careful pretesting is a must. Second, administration techniques should endeavor to make sure respondents are paying attention. Phone- and in-person enumerators can be trained to slow down or confirm responses to list experiment or other more complicated question forms. They can also make subjective judgments about respondents' levels of engagement in the survey. We can imagine several design strategies for electronically-administered surveys to slow users down and induce them to pay more attention. For example, survey interfaces could randomly move the text and responses to different points around the screen so as to force users to at least minimally adjust. Confirmation stages for certain responses could be introduced. Silly questions can be included to see if respondents are paying attention. Forcing respondents to pay attention may or may not increase truthfulness but it will likely increase the number of non-truthful responses that are amenable to modeling as ceiling and floor effects relative to the less tractable random error situation. All these interventions have costs, however, and will preclude asking more substantively interesting questions. Best practices in this area have yet to be developed.

Given that some error is unavoidable, we would like to convert any systematic biases in these errors into random noise, i.e., make trop. If error is due to respondents repeatedly

clicking or answering in the same way in an effort to rush through the survey then some immediate solutions present themselves. In an electronic interface the survey designer can randomize the order in which the possible responses are presented. For example, in the interface depicted in figure 1 we would like the radio buttons for the responses to be shuffled randomly. Other options include pull-down menus where the ordering of the values can be randomized across questions or requiring the respondent to type in a numerical value, returning an error if the person typed in a number that is not admissible. When a respondent is not paying attention these strategies have the virtue of converting what might be dangerous top- or bottom-biased error into something looking more like uniform error.

Ask both direct and indirect versions of the question. Echoing the advice in Blair and Imai (2012); Glynn (2013), asking both direct and indirect versions of the question is useful whenever doing so does not endanger respondents or enumerators. In addition to the reasons given by others, asking both versions of the question allows the analyst to compare whether those who answer affirmatively to the direct question are also those deemed likely to have the attribute when the list experiment is analyzed.

Ask calibration questions if possible. Both AMJ and Kiewiet de Jonge and Nickerson (2013) make good use of ancillary list experiment questions that have treatment items of either very low or very high prevalence in order to bound error rates and respondent performance. Asking such questions is costly, however, and may not we worthwhile in certain contexts. But they appear to be a useful tool to examining how the sample at hand is actually reacting to indirect questioning, especially at the extremes.

Consider multiple or other indirect question modes Rosenfeld, Imai and Shapiro (2015) conduct an exhaustive validation study of the three major types of indirect survey questions (list, endorsement, and randomized response). Consistent with results here they also find that the list experiment (analyzed with a Bayesian extension of the ICT-MLE approach) produces population prevalence estimates that are biased (relative to the known truth and

other question modes) yet still far better than direct questioning. Which question mode is most appropriate in a particular situation is not obvious and a multiple-method, triangulation strategy may be worth pursuing, although, again the costs in terms of time, cognitive demands on respondents, technical administration, and efficient use of data are all non-trivial.

Track and examine respondents' broader behavior in the survey when using electronically administered surveys. In determining the scale and type of respondent error researchers should get in the habit of tracking respondent behavior throughout the survey. Obviously this is easier an more accurate in computer-mediated modes. Some useful metrics include total survey spent on the survey, how long they spent on particular pages or questions, and whether they logged out and then completed the survey later.

The behavior of respondents at the extremes of the list experiment distribution should be of particular concern to researchers thinking of employing the ICT-MLE. Are these respondents spending less time on the list experiment page than other respondents? Are they answering nearby questions in a similar way? Is there straight-line behavior in other parts of the survey? If so is it systematically skewed in a particular direction?

Compare ICT-MLE and DiM estimates. Simple and transparent difference-in-means analysis should be the place to start. If covariates are not a concern in a particular application then ICT-MLE becomes even less attractive as an analysis tool. If ICT-MLE is used its results should be compared to those from DiM. If the underlying prevalence of the sensitive item is shown to be low and/or the two estimates diverge sharply this should be viewed as evidence that there is likely significant respondent error in the data. In interpreting this error analysts should obviously conduct the diagnostics for ceiling and floor effects described in Glynn (2013) and Blair and Imai (2012). Conditional on results from ceiling and floor analysis, large divergence between DiM and ICT-MLE, especially when DiM returns a null result, should be viewed as an indication that ICT-MLE results may not be reliable.

Care should be taken in using the ICT-MLE output as a covariate. The big selling point of the ICT-MLE is its ability to generate individual-level predictions that a particular respondent has the sensitive attribute. This individual-level ability is bought by invoking the individual-level no liars assumption. Imai, Park and Greene (2014) have taken the next logical step, building both two-stage and full likelihood models in which individual-level propensities to possess the sensitive attribute (estimated from ICT-MLE) are then used as predictors for another behavior of interest. For example, suppose a researcher runs a list experiment designed to ask respondents about racial attitudes toward African-Americans. ICT-MLE will yield estimates of each respondents level of anti-Black sentiment. The researcher might then want to use that quantity as a regressor in a model that predicts levels of support for President Obama.

The sensitivity of the ICT-MLE to measurement error may make this strategy problematic in actual applied situations. Even small levels of respondent error can induce severe bias as well as over confidence in results. Building this bias into a second stage model, whether estimated sequentially or jointly, seems hazardous. While formal Monte Carlo work incorporating measurement error for this specific enterprise remains to be done the results here should give pause.

5 Conclusion

Inspired by the odd findings in AMJ's real world application of the ICT-MLE, this paper further considered the dangers of random measurement error (as opposed to strategic misrepresentation) for the analysis of list experiments. We interrogated the key, individual-level *no liars* assumption needed to identify the ICT-MLE. This assumption requires that all responses in the extremes of the treatment-group distribution be viewed as truthful revelations of the respondent's status on the sensitive item. The conventional difference-in-means

estimator does not require the individual-level no liars assumption for unbiased estimation.

I argued that the no liars assumption is contrary to the applied researcher's rationale for using a list experiment in the first place, namely that respondents are reticent about truthfully revealing their status on the sensitive item. I showed that the ICT-MLE is particularly sensitive to deviations from this assumption both theoretically and computationally, especially when the number of respondents in the extremes is small. I then pointed out that current best practices for the design of list experiments entail minimizing the number of respondents appearing in the extremes of the response distribution–exactly the cells that the ICT-MLE requires for identification and estimation. In other words, compounding problems of measurement error, design best practices are directly at odds with the computational requirements of the ICT-MLE. The ICT-MLE relies heavily on the assumed absence of error in parts of the treatment group response distribution that are subject to very small samples and carry an elevated risk of being erroneous based on the applied situation motivating indirect questioning in the first place.

We then constructed a series of Monte Carlo experiments, comparing the difference-in-means estimator to the ICT-MLE under different rates and types of measurement error. We found that inducing error creates problems for both estimators, but the ICT-MLE is far more sensitive. Survey respondent error, even at low levels can induce severe bias in the ICT-MLE depending on the type of error and the underlying frequency of the sensitive item. The difference-in-means estimator, while not impervious to respondent error, is much less prone to generate erroneous inference when survey responses are measured with error.

Based on these findings I offered some preliminary advice for applied researchers. Key among them are pretest carefully and start with DiM analysis. When ICT-MLE and DiM estimates diverge be extremely cautious, especially when the sensitive item is found to be rare. The extent to which measurement error causes problems for either of Imai, Park and Greene (2014)'s two-stage or full likelihood models is an open question for future research.

Similarly, technical and survey mode interventions designed to mitigate measurement error problems in list experiments remain to be tested.

A Derivation of alternative DiM calculation

Letting $N_1^x = |\{i : y_i = x, T_i = 1\}|$ and $(N - N_1)^x = |\{i : y_i = x, T_i = 0\}|$ we can rewrite equation 3 as

$$\begin{aligned}
\hat{\tau} &= \frac{1}{N_1}\left[(J+1)N_1^{J+1} + JN_1^J + \ldots + 0N_1^0\right] - \frac{1}{N-N_1}\left[J(N-N_1)^J + \ldots + 0(N-N_1)^0\right] \\
&= \frac{1}{N_1}\sum_{j=0}^{J+1}\sum_{i=j}^{J+1} N_1^i - \frac{1}{N-N_1}\sum_{j=0}^{J}\sum_{i=j}^{J}(N-N_1)^i \\
&= \sum_{k=0}^{J+1}\left[\frac{1}{N_1}\sum_{j=k}^{J+1} N_1^j - \frac{1}{N-N_1}\sum_{j=k}^{J}(N-N_1)^j\right]
\end{aligned}$$

That is, we can calculate the DiM estimator by taking the difference between baseline proportion and treament proportion of respondents saying at least j and then summing these differences across $j = 0, \ldots, J+1$

B AMJ experiment questions

Table 4: Impossible event list experiment (September 2013 wave)

Prompt:	"Here are some things that may have happened to you during the past twelve months. HOW MANY of these events happened to you?"
1	"I was asked to serve on a jury"
2	"I was called by a telemarketer"
3	"I was audited by the IRS (Internal Revenue Service)"[15]
4	"An airline cancelled my flight reservation"
Treatment	"I was abducted by extraterrestrials (aliens from another planet)."

Table 5: Common illegal/undesirable behavior list experiment (September 2013 wave)

Prompt:	"Here are some things that you might have done during the past 30 days. HOW MANY did you do?"
1	"I travelled to a foreign country"
2	"I flossed my teeth"
3	"I littered in a public place"
4	"I celebrated my birthday"
Treatment	"I read or wrote a text (SMS) message while driving"

References

Ahlquist, John S., Kenneth R. Mayer and Simon Jackman. 2014. "Alien Abduction and Voter Impersonation in the 2012 US General Election: evidence from a survey list experiment." *Election Law Journal* 13(4):460–75.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444–455.

Blair, Graeme and Kosuke Imai. 2010. "list: Statistical Methods for the Item Count Technique and List Experiment." Available at The Comprehensive R Archive Network (CRAN).
URL: *http://CRAN.R-project.org/package=list*

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(4):1043–63.

Chaudhuri, A. and T.C. Christofides. 2007. "Item Count Technique in Estimating the Proportion of People with a Sensitive Feature." *Journal of Statistical Planning and Inference* 187:589–593.

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

Corstange, Daniel. 2012. "Vote Trafficking in Lebanon." *International Journal of Middle East Studies* 44(3):483–505.

Frye, Timothy, Ora John Reuter and David Szakonyi. 2014. "Political Machines at Work: Voter Mobilization and Electoral Subversion in the Workplace." *World Politics* 66(2).

Gelman, Andrew. 2014.
URL: *http://andrewgelman.com/2014/04/23/thinking-list-experiment-heres-list-reasons-think/*

Gingerich, D.W. 2010. "Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys." *Political Analysis* 18:349–80.

Glynn, Adam. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77:159–72.

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Melendez, Javier Osorio and David W. Nickerson. 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56(1).

Gooch, Andrew and Lynn Vavreck. forthcoming. "How Face-to-Face Interviews and Cognitive Skill Affect Item Non-Response: A Randomized Experiment Assigning Mode of Interview." *Poltical Science Research and Methods* .

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

Imai, Kosuke, Bethany Park and Kenneth F. Greene. 2014. "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models." *Political Analysis* .

Internal Revenue Service. 2014. "Internal Revenue Service Fiscal Year 2013 Enforcement and Service Results.".
URL: *http://www.irs.gov/PUP/newsroom/FY%202013%20Enforcement%20and%20Service%20Results%20WEB.pdf*

Kiewiet de Jonge, Chad P. and David W. Nickerson. 2013. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* pp. 1–24.
URL: *http://dx.doi.org/10.1007/s11109-013-9249-x*

King, Gary and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9:137–63.

Kuha, Jouni and Jonathan Jackson. 2014. "The item count method for sensitive survey questions: modelling criminal behaviour." *Journal of the Royal Statistical Society Series C: applied statistics* 63(2):321–41.

Kuklinski, J. H., M.D. Cobb and M. Gilens. 1997. "Racial Attitudes and the "New South"." *Journal of Politics* 59(2):323–49.

Madden, Mary and Lee Rainie. 2010. Adults and Cell Phone Distractions. Technical report Pew Internet and American Life Project Washington, D.C.: .
URL: *http://www.distraction.gov/download/research-pdf/Adults-Cellphone-Distractions.pdf*

Naumann, Rebecca B. 2011. *Morbidity and Mortality Weekly Report* 62(10):177–82.
URL: *http://www.cdc.gov/mmwr/pdf/wk/mm6210.pdf*

Rosenfeld, Bryn, Kosuke Imai and Jacob N. Shapiro. 2015. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60(3):783–802.