List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators

John S. Ahlquist*

Forthcoming, *Political Analysis*

July 24, 2017

Abstract

The Item Count Technique (ICT-MLE) regression model for survey list experiments depends on assumptions about responses at the extremes (choosing no or all items on the list). Existing list experiment best-practices aim to minimize strategic misrepresentation in ways that virtually guarantee that a tiny number of respondents appear in the extrema. Under such conditions both the "no liars" identification assumption and the computational strategy used to estimate the ICT-MLE become difficult to sustain. I report the results of Monte Carlo experiments examining the sensitivity of the ICT-MLE and simple difference-in-means estimators to survey design choices and small amounts of non-strategic respondent error. I show that, compared to the difference-in-means, the performance of the ICT-MLE depends on list design. Both estimators are sensitive to measurement error, but the problems are more severe for the ICT-MLE as a direct consequence of the no liars assumption. These problems become extreme as the number of treatment group respondents choosing all the items on the list decreases. I document that such problems can arise in real world applications, provide guidance for applied work, and suggest directions for further research.

With the advent of cheap and reliable Internet surveys, indirect questioning methods for sensitive topics are easier and cheaper to deploy than ever.[1] The survey list experiment is

*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, Tom Pepinsky, Margaret Roberts, Michael D. Ward, Yiqing Xu and, Alex Tahk for helpful conversations.

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

among the most commonly used of these tools. Along with increased interest in list experiments have come new design procedures and statistical estimators for list experiment data (Aronow et al., 2015; Blair and Imai, 2012; Corstange, 2009; Glynn, 2013; Imai, 2011; Liu et al., 2017; Tian et al., 2014). The item count technique regression models, particularly the maximum likelihood estimator (ICT-MLE), is the statistical innovation that has justifiably received the most attention (Imai, 2011).[2] This estimator aims to tell us more about the relationships between covariates and the sensitive behavior than traditional difference-in-means analysis (DiM). To achieve this the ICT-MLE leans on data in the extremes of the response distribution for the treatment group (answering 0 or giving the maximum number of items on the list). Consequently the ICT-MLE requires strong assumptions about the truthfulness of respondents' answers—Imai's "no liars" identification assumption—in a situation where we already doubt respondents' willingness to reveal their status.

This paper interrogates this key assumption in conjunction with current list experiment design best practices. I argue that following current list experiment design best-practices aimed at minimizing *strategic* misrepresentation implies that there will be a tiny number of responses in the extremes of the response distribution. The small numbers involved combined with the expectation that respondents want avoid revealing their status on the sensitive item imply that these few responses are particularly likely to have resulted from *non-strategic* measurement error, something completely ignored in the existing literature. The ICT-MLE, by construction, will be sensitive to small samples and therefore measurement or respondent error, especially when it appears in the extremes.

I recapitulate how assumptions regarding respondent accuracy are critical to the performance of the ICT-MLE. The DiM estimator requires weaker assumptions for for unbiasedness and is not directly affected by the number of responses in the extremes. I present results

[2]Imai (2011) also proposes nonlinear least squares estimator in addition to the MLE. The MLE has been the more widely used in applied work and forms the basis for several extensions (e.g., Eady (2017)). I defer evaluation of the NLS estimator for future work.

from a series of Monte Carlo experiments comparing the performance of ICT-MLE to simple difference-in-means analysis. Unlike earlier Monte Carlo studies (Blair and Imai, 2012), the simulations reported here incorporate design best-practices. When list design advice is followed, the ICT-MLE performs poorly even in the absence of measurement error due to its reliance on a tiny number of responses in the extremes. I then introduce non-strategic measurement error. These Monte Carlo results lead to further conclusions: 1) while respondent error induces bias in both the difference-in-means and ICT-ML estimators, the ICT-ML estimator is more sensitive, especially when this error appears in the top of the response distribution. 2) The problems respondent error induces become more severe as the number of truthful respondents in the extrema of the response distribution among the treated declines. This can result from *either* low prevalence of the sensitive item in the population or following survey design best-practices. 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 for strategic misrepresentation. I conclude by highlighting potential solutions that will not work and suggesting possible strategies for mitigating these problems and making research design trade-offs.

1 Extracting information from list experiments

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 indirect questioning only when she has reason to believe that:

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

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 sensitive item. These respondents are still forced to choose between either truthfully revealing their status or strategically misrepresenting their answers. To the extent these situations are present and respondents dissemble the list experiment is said to exhibit ceiling effects (i) and floor effects (ii). Ceiling and floor effects can be viewed as *strategic* measurement error in which some respondents who should appear in the extreme categories chose not to report those privacy-leaking values.

1.1 Difference in means estimator

Simple differences in means is the traditional analysis tool for list experiments. Formally, suppose we have 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 $\Pr(Z_i^* = 1 \mid X_i; \theta)$. 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 for π_{Z^*} 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

$$Y_i^* = \alpha + \tau T_i + \varepsilon_i \quad (4)$$

$$\varepsilon_i = e_i + \epsilon_i$$

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 $\mathrm{Cov}(T_i, \varepsilon_i) = 0$. The random assignment of T_i achieves this so long as measurement error is uncorrelated with T_i. The DiM estimator only requires two sums; measurement error can occur anywhere in the sum so long as 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.

1.2 The ICT-ML model

Glynn (2013) invokes stronger assumptions in order to characterize the joint distribution of $(Y_i(0), Z_i^*)$ and identify "joint proportions." Employing similar logic Imai (2011) goes on to develop a maximum likelihood estimator for the joint distribution $(Y_i(0), Z_i(t))$. The ICT-MLE is allows for the inclusion of covariates information and produces individual-level predicted probabilities that a respondent possesses the sensitive attribute. To achieve this the ICT-MLE relies on three identification assumptions:

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

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

Define $\mathcal{J}(t, y)$ as the set of respondents with values $(T_i, Y_i) = (t, y)$. To derive the likelihood Imai (2011) specified $g(\mathbf{x}, \delta) = \Pr(Z_i(1) = 1 \mid \mathbf{X}_i = \mathbf{x})$ 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 with parameters $\delta, \psi_\mathbf{z}$. The various combinations of $g(\cdot)$ and $h_z(\cdot)$ define the components of the likelihood. The no liars assumption directly determines which observations appear in which parts of the composite likelihood. For example, $g(\mathbf{x}, \delta)h_1(J; \mathbf{x}, \psi_\mathbf{1})$ describes the contribution to the likelihood of a treatment group respondent who answered "$J + 1$." Similarly, the $\mathcal{J}(1, 0)$ respondents appear in the $h_0(0; \mathbf{x}, \psi_\mathbf{0})(1 - g(\mathbf{x}, \delta))$ term of the likelihood. As a result the no liars assumption affects the composition of the $\mathcal{J}(1, y)$ for the remainder of the likelihood.

The computational strategy pursued in Blair and Imai (2010, 2012); Imai (2011); Imai, Park and Greene (2015) involves treating the $Z_i(1)$ as partially missing data and then deriving a complete data likelihood that can be maximized via the EM algorithm. The "observed" $Z_i(1)$ are those observations in $\mathcal{J}(1,0) \cup \mathcal{J}(1, J+1)$. The no liars assumptions requires that we believe that these observations are observed without error.

1.3 Strategic and non-strategic respondent error

The no liars assumption is critical for our ability to extract more information from list experiment data. Scholars using list experiments are sensitive to the assumption about truthfulness in responses at the extremes, as our discussion of ceiling and floor effects shows. Survey design best-practices have long recognized the potential for list experiments to "leak" privacy due to such strategic behavior. Ceiling effects imply a downward bias in $\hat{\pi}_{Z^*}$, regardless of the estimator. With this in mind, Kuklinski, Cobb and Gilens (1997) argue that an appropriately designed list experiment will aggressively seek to minimize the number of respondents forced to choose between answering truthfully and revealing their sensitive status. They recommend including an item on the control list common in the population (to get off the floor) as well as an item that is rare (to avoid bumping into the ceiling). 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. Blair and Imai (2012) reiterate all this advice.

Considerable effort has also gone into diagnosing and modeling possible strategic misrepresentation. Aronow et al. (2015); Blair and Imai (2012); Chaudhuri and Christofides (2007); Glynn (2013) develop diagnostic tests and modeling extensions for floor and ceiling effects. Kuha and Jackson (2014) extend and improve the ICT-MLE algorithm and variance estimation. Eady (2017) extends the ICT-MLE (relying on the same identifying assumptions) to explicitly model *who* is most likely dissembling.

Without detracting from the work on ceiling/floor effects, it is worth highlighting that almost all work aimed at testing and relaxing ICT assumptions has focused on strategic behavior by respondents, 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. The presence or absence of ceiling/floor effects tells us nothing about whether nonstrategic error is also a serious concern. More importantly, existing model-based fixes for ceiling and floor effects treat respondent error in an entirely asymmetric fashion: we worry that respondents strategically choose not to reveal an extreme value, generating erroneous values in other parts of the response distribution, yet we simultaneously maintain the *assumption* that all the observed responses in the extreme categories are error-free observations. Gingerich et al. (2016) go so far as to call this assumption "one-sided lying."

In short, the ICT-MLE relies on the assumption that we can treat observed survey responses in the extremes of the treatment group distribution as completely truthful. But this assumption flies in the face of the concerns about respondent privacy and 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 ICT-ML model can be prone to small sample sizes. Current list experiment 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 no liars assumption. Existing Monte Carlo evidence supporting the ICT-MLE does not incorporate either of these challenges.

1.4 Non-strategic error and consequences

It is uncontroversial to assert that measurement error is endemic in surveys. The real question surrounds the type of error and consequences for various estimators. Let's consider some

hypothetical processes giving rise to non-strategic measurement error. One possibility is *uniform error*: 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 her treatment status. Uniform error will be correlated with the treatment status in the list experiment 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.[3] The ICT-MLE will also be biased and inconsistent under uniform error because the distributional assumptions are incorrect. Uniform error 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, 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 assumption.

Many other error processes are obviously possible. For our purposes here we will focus on "top-biased error," a process by which the respondent's truthful response is randomly replaced with the maximum value available to her. 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 estimators.[4]

[3]To see the intuition 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 $\mathrm{Cov}(T_i, e_i) \to 0$.

[4]If we were to somehow correctly assume a particular error process then it might be modeled. Correctly assuming an error process seems very unlikely and any real population would almost surely exhibit a mixture of error processes, all invisible to the researcher. Trying to address nonstrategic error by making more and stronger assumptions seems like a high-cost, low-return strategy.

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 present serious problems for the ICT-MLE as the observed "$J+1$" responses are all treated as truthful. 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.

This simple discussion has two important implications that inform the Monte Carlo experiments below. First, whatever problems non-strategic error induces will be exacerbated as the number of responses in the extremes of the treatment group response distribution decreases. Very small samples in these extreme cells can occur either because the population prevalence of the sensitive item is low or because the survey is well-designed and few respondents actually fell in the extremes of the distribution. Since 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.

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 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. Again such problems will be exacerbated when sample sizes in the extrema are small.

2 Monte Carlo Experiments

I conduct a series of Monte Carlo experiments to better describe how the DiM and ICT-MLE respond to list experiment design, sensitive item prevalence, and respondent error. The Monte Carlo simulations represent a $3 \times 2 \times 2 \times 2$ design in which I vary the prevalence of the sensitive item (low, medium, and high), the length of the list ($J \in \{3, 4\}$), the design of the control item list, and the presence of top-biased error. In supplementary materials I also vary the size of the sample.[5]

I investigate three different levels of prevalence for the sensitive item. Under each scenario a "respondent," i, possesses the sensitive attribute with probability given by

$$\Pr(Z_i^k = 1) = \text{logit}^{-1}(b_0 + b_1^k x_i), \quad k \in \{L, M, H\} \tag{5}$$

The only covariate is $X \sim U[0, 1]$. Following the Monte Carlo simulations in Blair and Imai (2012) I fix $b_0 = 0$ and $b_1^H = 1$. I set $b_1^M = -2$ and $b_1^L = -4$, implying that the underlying population prevalences are $\pi_{Z^*}^L \approx 0.12$, $\pi_{Z^*}^M \approx 0.27$, and $\pi_{Z^*}^H \approx 0.62$.

In manipulating the list design I investigate two different sets of list structures. The first set of lists generates control items following protocols Blair and Imai (2012) borrow from Corstange (2009). In these lists all control items are independent. In the second set—referred to as the "designed" lists—I construct the control item lists to conform to current recommendations for avoiding strategic misrepresentation. There are high- and low-prevalence control items; in the $J = 4$ list there is also negative correlation between two of the control items; further details are in the supplementary materials and related `R` code.

For each of the 2000 Monte Carlo runs I generate a sample of $N = 1000$ "respondents."[6]

[5] All data and code as well as the output and logs of the actual Monte Carlo simulations are available on the dataverse repository associated with this article: doi:10.7910/DVN/ELHTWJ.

[6] 1000 respondents is a common sample size for list experiments. Rosenfeld, Imai and Shapiro (2015) use $N = 1352$ to evaluate list experiments in their validation study. I also ran the same experiments with $N = 2000$. Increasing the sample size has no material consequences for the conclusions generated here.

With equal probability I randomly assign each of the respondents 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 $k \in \{L, M, H\}$. This value represents the data we would hope to observe in list experiments satisfying Imai's three basic identification assumptions with no measurement error but under different structures of the control lists and different population frequencies for the sensitive item. Table 1 displays the number of respondents we expect to appear in $\mathcal{J}(1, J+1)$ under the two list constructions. Designed lists sharply reduce the number of respondents in the top category relative to the Blair-Imai construction.

Table 1: Expected number of respondents in $\mathcal{J}(1, J+1)$ under the different list scenarios for $N = 1000$, equal probability of treatment assignment, and no error.

	Blair-Imai		Designed	
π_{Z^*}	$J=3$	$J=4$	$J=3$	$J=4$
High	38.9	11.5	4.7	4.0
Mid	16.8	5.0	2.0	1.7
Low	7.5	2.2	0.9	0.8

I then introduce "top-biased" error by randomly selecting three percent of respondents. For each selected respondent, if $T_i = 0$ we replace y_i^k with J. If $T_i = 1$ we do the same thing only with a $J + 1$. Note that while measurement error induces a violation of the no liars assumption, it still leaves us with data that satisfy the randomization and no design effects assumptions. Thus we can view this study as examining the estimators' sensitivity to violations of the no liars condition.

Within each Monte Carlo iteration we fit two models to both the data with and without error. The first model is the ICT-MLE. Following Blair and Imai (2012) I estimate the default double binomial ICT-MLE and constrain $h_0(\mathbf{x}, \psi_0) = h_1(\mathbf{x}, \psi_1)$ with X as covariate. The

Details are reported in the supplementary materials.

second is the DiM, calculated as the OLS regression of Y_k on the treatment indicator with no covariate. I calculate heteroskedasticity-corrected standard errors for the OLS regression.

Beyond concerns with respondent error the Monte Carlo experiments here differ from existing simulation studies of the ICT-MLE in two ways. First, Blair and Imai (2012), following Corstange (2009), did not vary the prevalence of the sensitive item. They only considered what we are calling the "high prevalence" condition here. Second, the Blair-Imai study (again, following following Corstange (2009)) did not incorporate list experiment design practices meant to minimize strategic misrepresentation. These two differences have consequences for the number of respondents in the extreme cells of the distribution, explaining why my error-free results differ from theirs.

I evaluate the simulations on the following dimensions:

- Computational stability,
- Bias, variance, and coverage in covariate parameter estimates among the ICT-MLE,
- Bias, variance, and coverage in the population prevalence estimate for the sensitive item, from the ICT-MLE and DiM.

2.1 Results

2.1.1 Computational stability

We first consider the computational stability of the ICT-MLE; the DiM estimator had no computational difficulties. I define instability as the algorithm exiting with an error.[7]

Recall that the ICT-MLE uses 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. The lower the prevalence of the sensitive

[7]This was typically due to computationally singular matricies during the M-step of EM.

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 $\mathcal{J}(1, J+1)$. We therefore expect the ICT-MLE to be most unstable in the low prevalence, no error condition with the designed list.

Among the twelve Blair-Imai lists there were no stability problems. In those scenarios the mean number of $\mathcal{J}(1, J+1)$ responses ranged from 3.1 (in the $J = 4$, low prevalence, no error list) to 52.5 (in the $J = 3$, high prevalence, error list). I only observe 3% of runs with $\mathcal{J}(1, J+1)$ empty–all in the $J = 4$, low prevalence, no error condition.

Figure 1 displays stability results for the twelve designed list experiments, where we see different behavior. The ICT-MLE becomes increasingly fragile as the number of responses in $\mathcal{J}(1, J+1)$ declines. In the low-prevalence $J = 4$ condition the algorithm failed over five percent of the time. This rate declines as the number of observations in $\mathcal{J}(1, J+1)$ increases. The figure also confirms that inducing error has a similar effect on the stability of the algorithm; error at the top extreme, even at low levels, prevented computational failure almost entirely. This has some problematic implications: when the list is designed to minimize strategic behavior and 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 can be misleading (as we shall see).

2.1.2 Bias, variance, and coverage for regression parameters

I begin with the results for the regression parameter in part because a selling point of the ICT regression framework is the ability to include covariates. But more importantly the results for the regression parameter will help interpret ICT-MLE simulation results for population quantities. Focusing on b_1, Figure 2 displays bias (2a), root-mean-square-error (RMSE; 2b), and coverage results (2c) as returned under the Blair-Imai lists. Consistent with the results in Blair and Imai (2012), the estimator is unbiased at high frequencies for the sensitive item,

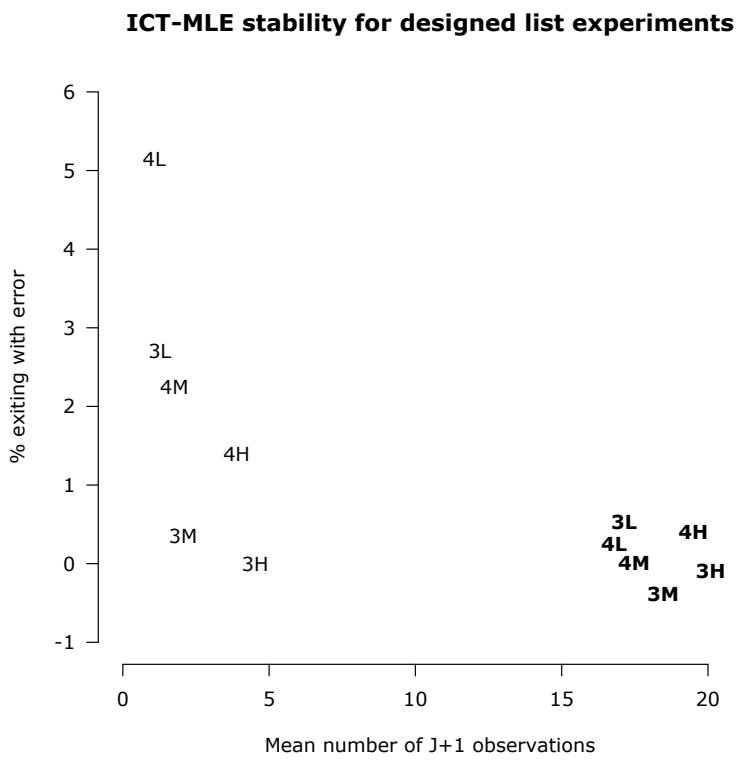
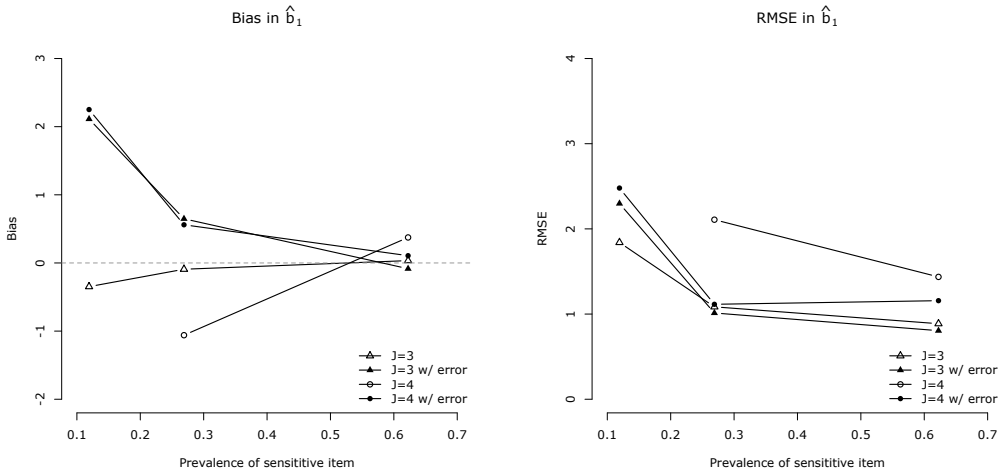


Figure 1: Percent of Monte Carlo runs in which the ICT-MLE exited with an error as a function of the mean number of observations in $\mathcal{J}(1, J+1)$. Text indicates the list condition, e.g., 3M refers to the $J = 3$ list with medium prevalence for the sensitive item. Bold-text items are from runs with 3% top-biased error and jittered for clarity.

even with error included. The $J = 3$ list is, as expected, lower variance than the $J = 4$ list. The ICT-MLE under the $J = 4$ list starts to degrade at moderate levels of π_{Z^*} even without error due to its sensitivity to small changes in the number of responses in $\mathcal{J}(1, J+1)$. In the low prevalence condition with no error the ICT-MLE is unstable, so I do not include bias and RMSE quantities in the plot for this case.[8]

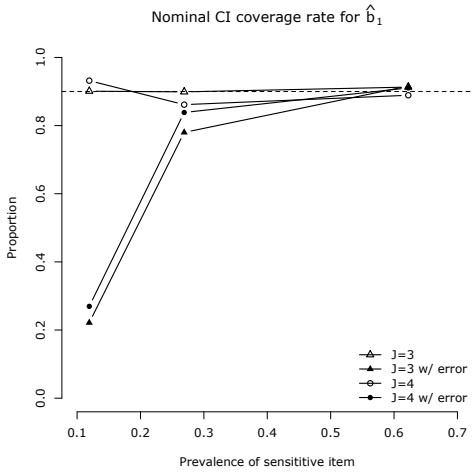
Adding the 3% error has the two anticipated effects. First, it stabilizes the ICT-MLE by creating more (erroneous) observations in the $\mathcal{J}(1, J+1)$ category, mitigating the rare events problem in the low-prevalence condition. Second, it induces bias at moderate and,

[8]ICT-MLE returned point estimates for b_1^L between -25408 and 14.5 with a mean -42 of for $J = 4$. This compares to a range of [-2.6,5] for the $J = 3$ high prevalence, no error case.



(a) Bias

(b) RMSE



(c) Coverage

Figure 2: Bias, RMSE, and 90% confidence interval coverage rates in $\hat{b}_1$ for the ICT-MLE applied to the Blair & Imai-style list experiments. Bias and RMSE for the $J = 4$, low-prevalence, no-error conditions are -38 and 903, respectively. These values are omitted from Figures 2a and 2b for clarity in presentation.

especially, low frequencies for the sensitive item. The ICT-ML estimate for b_1 is sensitive to violations of the no liars assumption even when the population prevalence is moderate and when the list is not specifically designed to minimize strategic behavior. The bottom panel displays nominal 90% confidence interval coverage, further confirming expectations. Not only does a small amount of error induce bias in the ICT-MLE but it also raises the risk

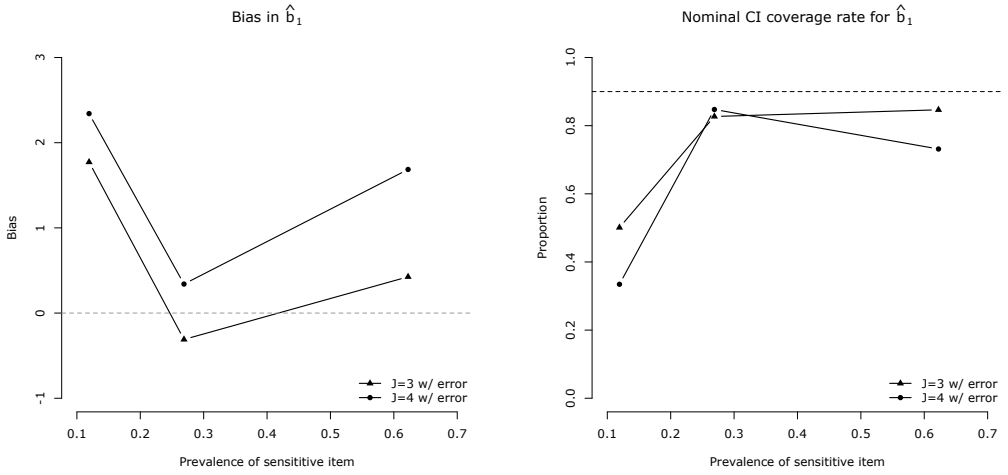
of erroneous inference.

Given the findings for the Blair-Imai list it should come as no surprise that stability and performance under the designed list is poor. Table 2 summarizes the distributions of the b_1 point estimates for each of the designed lists. Only in the high-prevalence situations did the ICT-MLE produce a stable distribution of estimates and only for the $J = 3$ list was the estimate approximately unbiased. The distance between the median and mean for the remainder gives some idea as to the instability of the estimates; this gap widens as the number of observations in the $\mathcal{J}(1, J + 1)$ declines. With such instability, concerns about coverage seem misplaced.

Table 2: The distributions of point estimates for b_1 from the ICT-MLE applied to the "designed" lists with no error.

List	1st Qu.	Median	Mean	3rd Qu.
$J = 3$, high	0.4	1.2	1.2	1.9
$J = 3$, mid	-7.5	-5.7	-28.5	-3.9
$J = 3$, low	-14.0	-9.4	-218.6	-6.5
$J = 4$, high	0.7	3.1	4.0	6.2
$J = 4$, mid	-8.4	-4.7	-523.9	-2.3
$J = 4$, low	-17.5	-9.4	-622.1	-5.6

As with the Blair-Imai lists, the introduction of error has the perverse effect of stabilizing the ICT-MLE for the designed lists as there are more observations in the extreme categories. Figure 3 displays bias (3a) and coverage rates (3b) for $\hat{b}_1$ from the designed list experiments with 3% top-biased error. Unfortunately, the computational stabilization comes around biased estimates of the regression parameter; bias is worst when the sensitive item is rare. Looking at panel 3b we see that the ICT-MLE generates standard errors that are systematically too narrow, dramatically so in the low prevalence condition. But while problems are most severe when the sensitive item is rare, they are not necessarily restricted to such situations. The results for the designed lists show that the performance of the ICT-MLE depends on decisions about the construction of the control list items.



(a) Bias (b) Coverage

Figure 3: Bias and 90% confidence interval coverage rates in $\hat{b}_1$ for the ICT-MLE applied to the "designed" list experiments with 3% top-biased error, as a function of list length and the prevalence of the sensitive item.

2.1.3 Bias, variance, and coverage in population prevalence estimates

In examining the population prevalence estimates we compare DiM and ICT-MLE. Figures 4, 5, and 6 present bias, RMSE, and 90% confidence interval coverage results, respectively, from the Blair and Imai-based lists. The solid line represents the ICT-MLE estimations while the broken line is the simple difference-in-means.

Mirroring the simulation results in Imai (2011), Figure 4 shows that the ICT-MLE and DiM are both unbiased estimators of π_{Z^*} when the population prevalence is high, even in the presence of measurement error. We do, however, continue to see ICT-MLE instability in the $J = 4$ list under no error, unsurprising given the regression parameter instability seen above. As the sensitive item becomes rare we see some apparent degradation in both estimators. But the introduction of small amounts of error causes problems. As expected, both estimators become more biased but the ICT-MLE suffers more at both the mid and low prevalence levels. At low prevalence the ICT-MLE is over-estimating the prevalence of the sensitive item by about 15 percentage points; the DiM estimator's bias is less than half

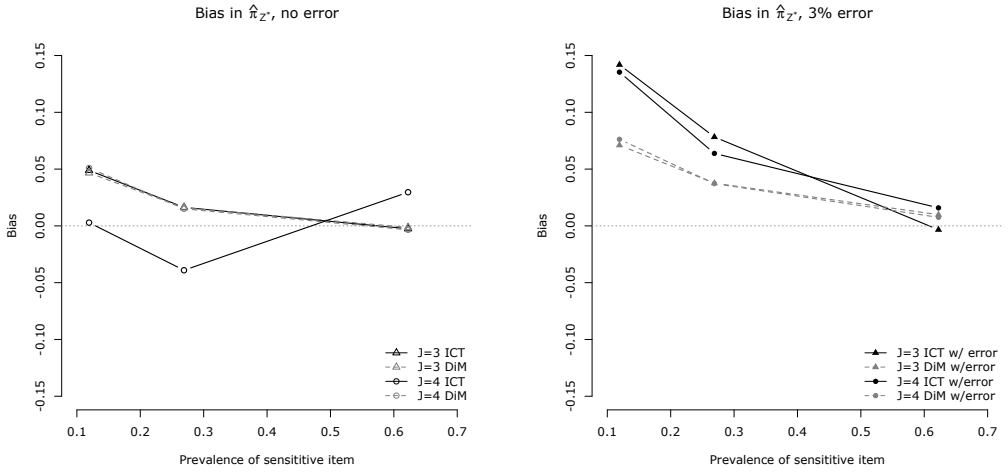


Figure 4: Bias in $\hat{\pi}_{Z^*}$ for the ICT-MLE (solid) and DiM (broken line) applied to the Blair-Imai style list experiments as a function of list length, the prevalence of the sensitive item, and presence of 3% top-biased error.

as bad.

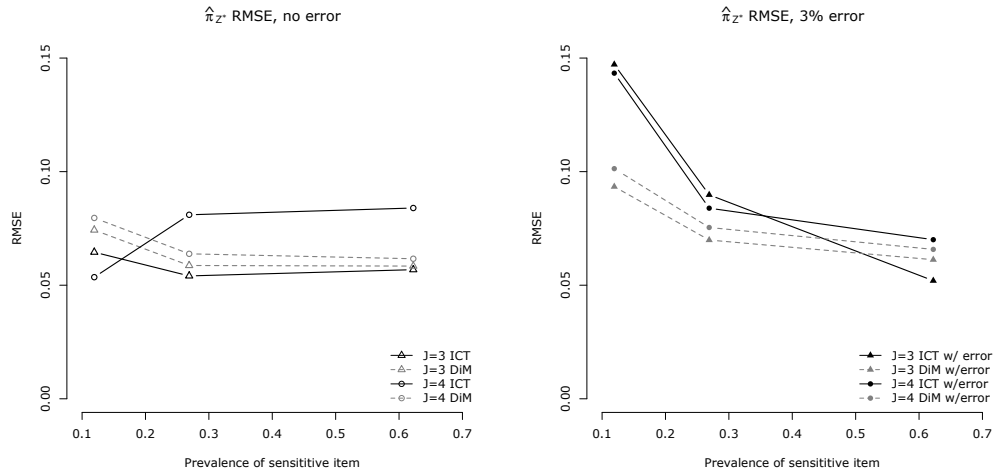


Figure 5: RMSE for $\hat{\pi}_{Z^*}$ for the ICT-MLE (solid) and DiM (broken line) applied to the Blair-Imai style list experiments as a function of list length, the prevalence of the sensitive item, and presence of 3% top-biased error.

Similar conclusions obtain from the RMSE results in Figure 5. The ICT-MLE with $J = 4$ list shows more variability when there is no error present. But the ICT-MLE is again more sensitive to measurement error than the DiM estimator. Figure 6 shows concerns with CI

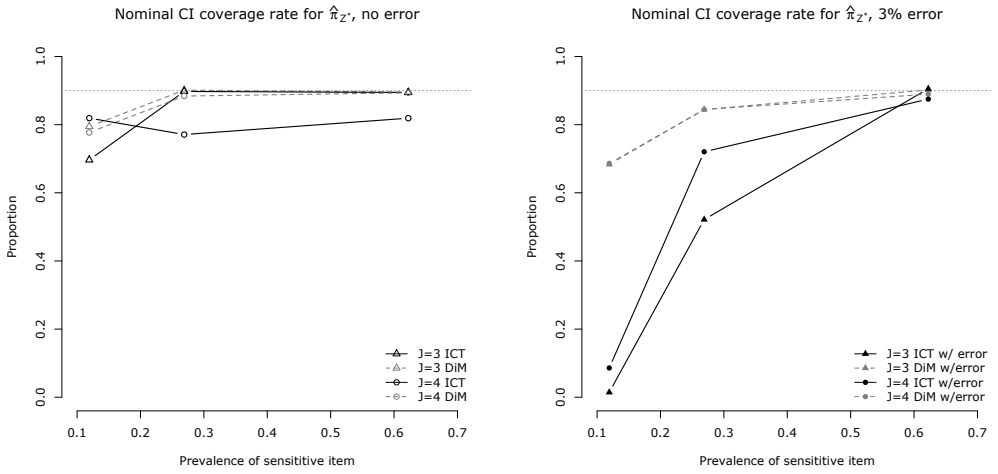


Figure 6: 90% confidence interval coverage rates for $\hat{\pi}_{Z^*}$ for the ICT-MLE (solid) and DiM (broken line) applied to the Blair-Imai style list experiments as a function of list length, the prevalence of the sensitive item, and presence of 3% top-biased error.

coverage and inference are again borne out when estimating population prevalence. With small amounts of measurement error the ICT-MLE returns standard error estimates that are too narrow when the sensitive item occurs at both moderate and low frequencies. The DiM results also degrade but to a far lesser extent.

Given the findings for the Blair-Imai lists, we should again expect poor performance under the designed list. Figure 7 presents bias and variance results from the "designed" list experiments in somewhat different format, reflecting the extreme variability in some of the ICT-MLE results. For example, in the $J = 4$ high prevalence lists without error (black broken lines) we observe π_{Z^*} estimates across the entire [0,1] interval. Even without error the ICT-MLE is giving biased and variable results under the designed list in all cases. The DiM estimator is stable throughout and unbiased at high prevalence both with and without error. At moderate prevalence we begin to see error causing some over-estimation under the DiM. At low prevalence we see some bias in the DiM without error, worsening once error is introduced. For example, in the $J = 4$ low prevalence condition the DiM shows bias of about 0.05 without error, worsening to 0.07 with error. In all cases the introduction of error

increases the variability of the DiM, as we would expect.

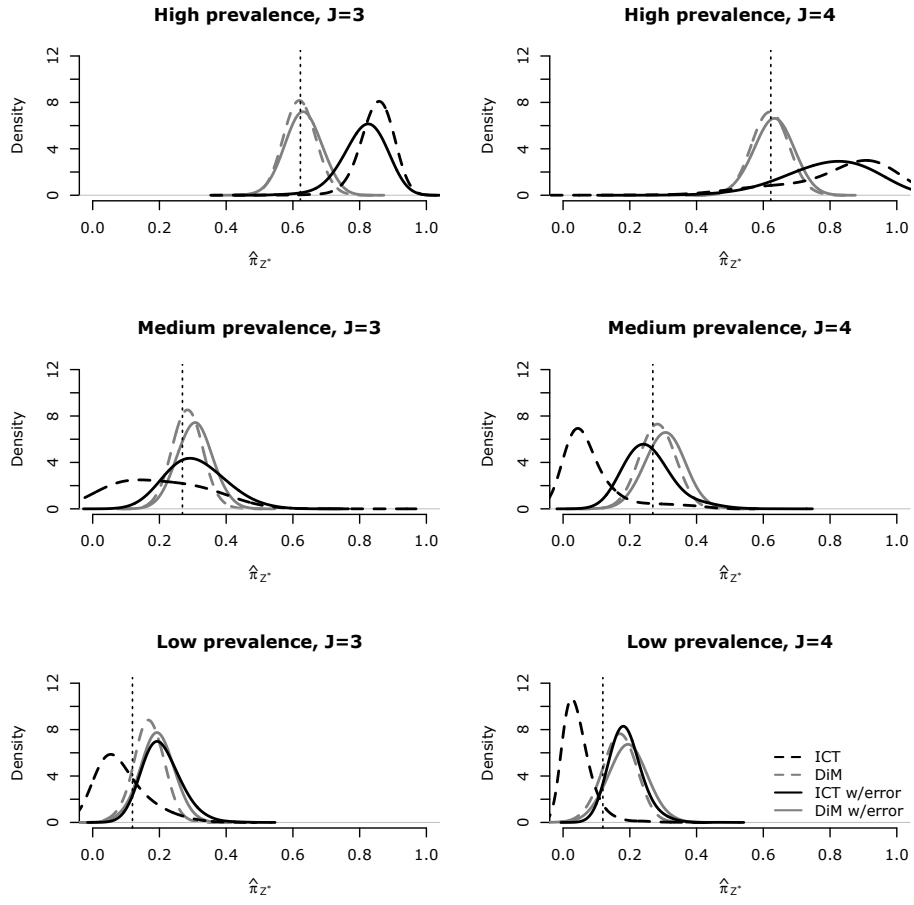


Figure 7: The distribution of point estimates of π_{Z^*} from the ICT-ML (black) and difference-in-means (grey) estimators applied to the "designed" list experiments, as a function of list length (left v. right), the prevalence of the sensitive item, and presence of 3% top-biased error (solid v. broken lines).

As before, adding error has the perverse effect of stabilizing the ICT-MLE, reducing both bias and variance in the low- and medium prevalence cases for both $J = 3$ and $J = 4$. This a consequences of the fact that under the no error condition the number of runs with few (or no) observations in $\mathcal{J}(1, J + 1)$ pulled the ICT-MLE toward a population estimate of 0. Adding error reduces those problems in the high- and moderate prevalence cases, with the apparent reduction in bias. But the same mechanism has the opposite effect as low

prevalence: a significant downward bias without error becomes a significant *upward* bias with error. This is most clear in the $J = 4$ condition where the ICT-MLE significantly *under* estimates π_{Z^*} on average but, once error is introduced, we see a significant overestimate.

For completeness, Figure 8 displays coverage rates for nominal 90% confidence intervals for $\hat{\pi}_{Z^*}$. Both with and without error the DiM estimator is stable and performing exactly as it did under the Blair-Imai lists. The ICT-MLE is unstable and has potential to yield misleading estimates under the designed lists regardless of whether we include error.

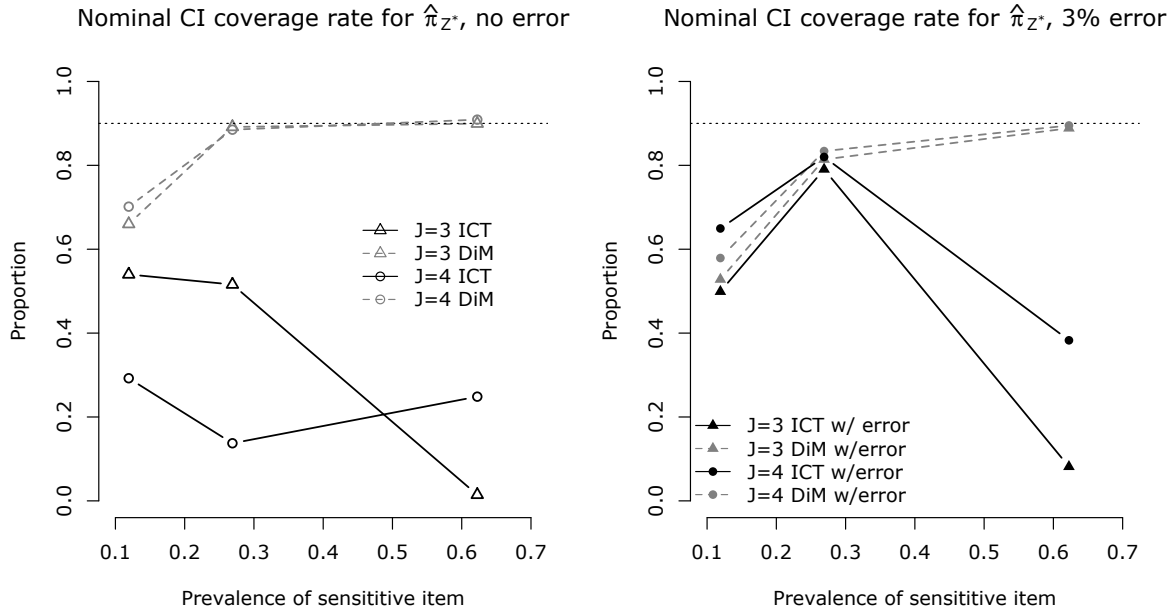


Figure 8: 90% confidence interval coverage rates for $\hat{\pi}_{Z^*}$ from the ICT-ML (black) and difference-in-means (grey) estimators applied to the "designed" list experiments, as a function of list length, the prevalence of the sensitive item, and presence of 3% top-biased error (left v. right).

The Monte Carlo simulations demonstrate that the performance of the ICT-MLE is sensitive to the design of the underlying control-item lists in ways that the DiM estimator is not. Both estimators are sensitive to rare sensitive items and error, but the ICT-MLE more so.

3 An example

To demonstrate that this problem of small samples in the extremes, 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 in the 2012 US election. AMJ find no evidence for substantial rates of voter impersonation but there are some anomalies. First, the DiM estimates and ICT-MLE estimates differ noticeably, with the ICT-MLE showing point estimates substantially larger than those using the simple mean comparison procedure. Second, some of the respondents (about 2.5% of the treatment sample, twelve individuals) in the voter impersonation question claimed the maximum number of items (five). If we maintain the no liars assumption then these twelve respondents are admitting to voter impersonation (in addition to a variety of other things). 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.

Examining the broader survey behavior of the respondents who claimed the maximum of five, AMJ find additional reason to treat these responses as suspect. For example 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. To further investigate this conjecture of respondent error AMJ fielded a second set of list experiments in September 2013 with a new YouGov sample and $N = 3000$, 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.

In addition to replicating the original list experiment questions AMJ fielded two more list experiments as calibration exercises. The first new question offered subjects the oppor-

tunity to admit to something believed to occur with (near?)zero probability: abduction by extraterrestrials. 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. 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 while Naumann (2011) estimates that about 31% of U.S. drivers aged 18-64 had sent an SMS while driving in the last 30 days. Both surveys used direct questioning techniques. The details of these lists experiments are described in supplementary materials. Both are $J = 4$ lists and both attempt to include both high-and low-frequency as well as plausibly negatively correlated items on the control list.

Figure 9 displays population prevalence estimates for all three list experiments as calculated using both DiM and ICT regression.[9] Several things are immediately apparent. First, 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 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 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 were they to be taken at face value.

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 3 displays their findings.

[9]ICT models use the double-binomial maximum likelihood estimator and ignore survey weights reported in the original paper. In fitting the ICT regressions we included age, race, and gender as covariates. Models passed tests for ceiling and floor effects.

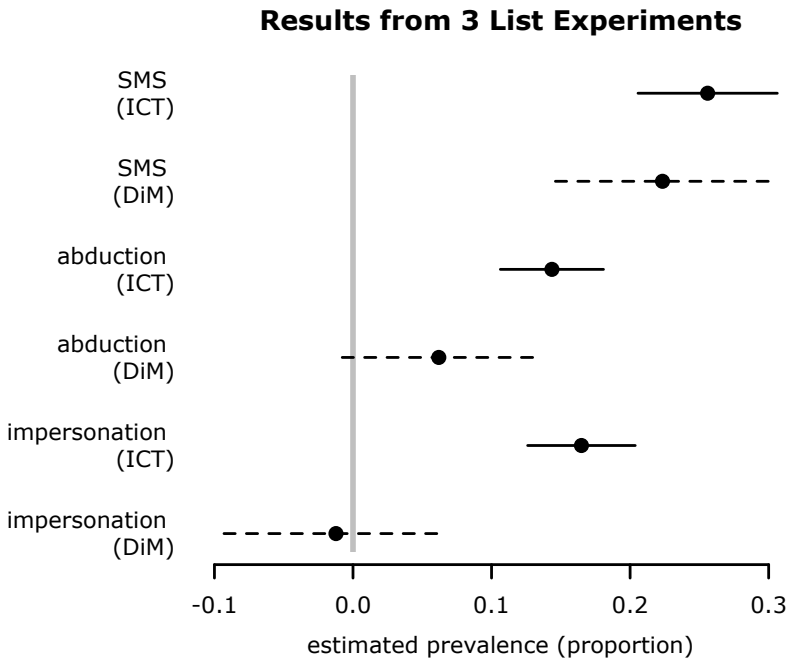


Figure 9: Evidence of problems with the ICT-MLE across three list experiments. Bars are 95% CIs for the DiM estimates and ±2SEs for the ICT estimates.

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. Maintaining the no liars assumption in that context corresponds to believing that all of these 36 respondents were abducted by aliens (and returned to answer the survey) *and* served on a jury *and* were audited by the IRS *and* had a flight canceled *and* received telemarketing calls. All in the same (unlucky) year. 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 probably 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. The respondents answering "5" in the

Table 3: The proportion of respondents selecting the most extreme value is 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

treatment conditions are both small in number and composed almost entirely of error. Second, in this situation the ICT-MLE overestimates the prevalence of two sensitive attributes, both of which have low (0?) population prevalence.

4 Implications and working toward solutions

The Monte Carlo simulations show that ICT-MLE is more sensitive to list design than the DiM, even in the absence of error. Unsurprisingly the addition of error can cause problems for both the DiM and ICT-MLE. The DiM, which does not require the no liars assumption, is the more robust in all cases. These findings lead us to consider—and reject—some possible solutions to the problem and then to consider when non-strategic error is more likely to arise. We conclude with some advice to applied researchers considering list experiments.

4.1 Strategies that will *not* work

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 case of the Wu-Hausman test. Unfortunately, the Monte Carlo results show that, depending

on the structure of the list and the prevalence of the sensitive item, the ICT-MLE may not generate 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.

The simulations and the example above show that small samples in the maximum values of the treatment distribution is serious problem for the ICT-MLE. It stands to reason that larger overall samples might mitigate this problem. But, for a fixed list experiment design and sampling frame, the expected number of responses in the top of the distribution will only grow linearly in N. Our designed Monte Carlo with $J = 3$ and high prevalence still averaged only 4.5 respondents in the top category of the treatment distribution. We would need a sample at least four times as large to begin to achieve some stability, and this is before any consideration of respondent error. The second wave of AMJ list experiments used $N = 3000$ and still ran in to problems. Their experience suggests that non-strategic respondent error happens at a relatively constant rate, something not solved by increasing N.

More generally ICT-MLE faces a conceptual difficulty: we turn to list experiments when we are worried that people do not want to reveal their status yet the ICT-MLE requires that we view anyone who does end up reporting a privacy-leaking value as a truthful revelation. Moreover, we actively try to ensure that nobody is put in a position whereby they are forced to choose between dissembling and reveling their status yet ICT-MLE estimation is based on these few respondents. Increasing the overall sample size in the hopes of ramping up the size of $\mathcal{J}(1, J + 1)$ fails to resolve this conflict.

4.2 Some advice

The findings presented thus far show that both small sample sizes in the extremes and non-strategic respondent error are problematic for list experiments but especially the ICT-ML estimator. We do not yet have tools for determining the levels, rates, and structure of non-strategic respondent error and developing such tools seems unlikely. But we can offer some

advice that builds on existing recommendations and tools.

4.2.1 Research and survey design

List experiments are weak tools for precise estimation of 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 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 estimator (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 is a step too far. If we already 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 too rare in the population for the survey to detect. Applied researchers must recognize that available tools are fragile in such situations.

Small samples in the extremes are particularly likely when we follow existing survey design advice for minimizing strategic misrepresentation and when the list is longer. There appears to be something of a trade-off: design a list experiment that will likely allow for more responses in the extremes and consider the ICT-MLE but risk more strategic misrepresentation or design a survey to minimize the chances strategic misrepresentation but rely on simpler analysis tools. Whether the first choice is attractive will depend on the underlying

prevalence of the sensitive item and the researcher's confidence in her ability to model ceiling and floor effects, something we have not investigated here. Unfortunately these are *ex ante* decisions affecting survey design.

Echoing advice from several quarters, *ask direct and indirect versions of the question whenever possible.*[10] Several versions of combined direct and indirect questioning are possible, but asking respondents both the direct and indirect versions of the question yield the most information. Using this design, Aronow et al. (2015) exploit comparisons between list experiments and direct questioning to develop a placebo test that jointly tests all three of Imai's identification assumptions. This test's importance is amplified given the difficulty of characterizing non-strategic error. The comparison between those who answered truthfully in the direct question and reported a value of $J+1$ might give some information about "compliers" which could then be used to generate a more precise lower bound on π_{Z^*}. Comparing those who answer negatively to the direct question but $J+1$ in the list experiment enables us to put an approximate lower bound on the rates of non-strategic error. But, again, such numbers are likely to be quite small in an experiment designed to protect privacy.

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 low- or 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 for examining how the sample at hand is actually reacting to indirect questioning.

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

[10]Eady (2017) and Gingerich et al. (2016) develop tools that rely on asking both direct and indirect questions to extract more information about who dissembles and whether the additional variance of indirect questioning is worth the cost. Both sets of tools rely on the no liars assumption and their sensitivity to non-strategic error is an area for future work.

(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 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 the sample are all non-trivial.

4.2.2 Survey Implementation

Adjust survey administration to minimize non-strategic error. 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 non-strategic error situation. All these interventions have costs and best-practices in this area have yet to be developed.

Given that some error is unavoidable, we would like to convert any systematic (e.g. top-) biases in these errors into something less damaging. 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, radio buttons for the eligible responses can be shuffled randomly or moved around the screen. 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.

Track and examine respondents' broader behavior in the survey. Tracking respondent behavior throughout the survey can be useful in determining the scale and type of respondent error. Obviously this is easier and more accurate in computer-mediated modes where some useful metrics include total time 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?

All the forgoing items give further weight to the standard dictum that *careful pretesting is a must.* Pre-testing control list items seems particularly worthwhile since we often do not know whether particular items are negatively correlated, etc. Pretesting allows a researcher to formulate expectations about the number of observations that might appear in $\mathcal{J}(1, J+1)$ and adjust the list or the analysis strategy accordingly.

4.2.3 Diagnostics and Analysis

Examine the number of responses in the extremes of the treatment group distribution. Independent of any concerns with respondent error the Monte Carlos reported here show that the ICT-MLE performs poorly when the number of responses in the maximum category falls below about 20-25. If small samples in the extremes appear in a particular application then the ICT-MLE is a poor choice of analysis tool, especially if DiM analysis suggests that the sensitive item is rare.

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 (2015) 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 re-

searcher 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 list design and measurement error may make this strategy problematic in actual applied situations. Even small levels of respondent error can induce 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 list design and measurement error for this specific enterprise remains to be done the results here should give pause.

5 Conclusion

This paper considered the role of list design and the dangers of non-strategic measurement error (as opposed to strategic misrepresentation) for the analysis of survey list experiments. We interrogated the individual-level "no liars" assumption needed to identify the ICT-MLE and underpinning its numerous extensions. 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 of population prevalence.

I argued that the the no liars assumption is contrary to the applied researcher's rationale for using a list experiment, namely that respondents are reticent about truthfully revealing their status on the sensitive item; privacy-leaking responses should be treated skeptically, not credulously. Moreover, reducing the risk of strategic misrepresentation entails minimizing the number of respondents appearing in the extremes of the response distribution—exactly the cells that the ICT-MLE relies on for identification and estimation. These small samples are particularly prone to be the result of simple, non-strategic error.

Based on a series of Monte Carlo experiments I found that ICT-MLE is sensitive to list design regardless of error, unlike the DiM. Inducing non-strategic error creates problems for both estimators, but the ICT-MLE is more sensitive. Even small deviations from the no liars assumption can induce bias and other problems in the ICT-MLE. The extent of these problems depend on the structure of the control list and the underlying frequency of the sensitive item. The difference-in-means estimator, while not impervious to respondent error, is computationally stable and 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. The extent to which non-strategic error causes problems for either of Imai, Park and Greene (2015)'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 and other forms of indirect questioning remain to be developed and tested rigorously.

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.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford and Donald P. Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3(1):43–66.

Blair, Graeme and Kosuke Imai. 2010. "list: Statistical Methods for the Item Count Technique and List Experiment." Available at The Comprehensive R Archive Network.

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.

Eady, Gregory. 2017. "The Statistical Analysis of Misreporting on Sensitive Survey Questions." *Political Analysis* pp. 1–19.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho and Mauricio Ruiz-Vega. 2016. "When to Protect? Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior." *Political Analysis* 24:132–56.

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

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. 2015. "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models." *Political Analysis* 23:180–96.

Internal Revenue Service. 2014. "Internal Revenue Service Fiscal Year 2013 Enforcement and Service Results.".

Kiewiet de Jonge, Chad P. and David W. Nickerson. 2013. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* 36(3):1–24.

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.

Liu, Yin, Guo-Liang Tian, Qin Wu and Man-Lai Tang. 2017. "Poisson–Poisson item count techniques for surveys with sensitive discrete quantitative data." *Statistical Papers* .

Madden, Mary and Lee Rainie. 2010. Adults and Cell Phone Distractions. Technical report Pew Internet and American Life Project Washington, D.C.: .

Naumann, Rebecca B. 2011. *Morbidity and Mortality Weekly Report* 62(10):177–82.

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.

Tian, Guo-Liang, Man-Lai Tang, Qin Wu and Yin Liu. 2014. "Poisson and negative binomial item count techniques for surveys with sensitive question." *Statistical Methods in Medical Research* .