

COMBINING LIST EXPERIMENT AND DIRECT QUESTION ESTIMATES OF SENSITIVE BEHAVIOR PREVALENCE

PETER M. ARONOW*
ALEXANDER COPPOCK
FORREST W. CRAWFORD
DONALD P. GREEN

Survey respondents may give untruthful answers to sensitive questions when asked directly. In recent years, researchers have turned to the list experiment (also known as the item count technique) to overcome this difficulty. While list experiments are arguably less prone to bias than direct questioning, list experiments are also more susceptible to sampling variability. We show that researchers need not abandon direct questioning altogether in order to gain the advantages of list experimentation. We develop a nonparametric estimator of the prevalence of sensitive behaviors that combines list experimentation and direct questioning. We prove that this estimator is asymptotically more efficient than the standard difference-in-means estimator, and we provide a basis for inference using Wald-type confidence intervals. Additionally, leveraging information from the direct questioning, we derive two nonparametric placebo tests for assessing identifying assumptions underlying list

PETER M. ARONOW is Assistant Professor in the Department of Political Science, Yale University, 77 Prospect Street, New Haven, CT 06520, USA. ALEXANDER COPPOCK is a doctoral student in the Department of Political Science, Columbia University, 420 W. 118th Street, New York, NY 10027, USA. FORREST W. CRAWFORD is Assistant Professor in the Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA. DONALD P. GREEN is Professor of Political Science at Columbia University, 420 W. 118th Street, New York, NY 10027, USA. The authors are grateful to Columbia University, which funded components of this research but bears no responsibility for the content of this report. This research was reviewed and approved by the Institutional Review Board of Columbia University (IRB-AAAL2659). This work was supported by the National Center for Advancing Translational Science (NCATS) [CTSA Grant Number UL1 TR000142 and KL2 TR000140 to F.W.C.], components of the National Institutes of Health (NIH), and the NIH roadmap for Medical Research. Helpful comments from Xiaoxuan Cai, Albert Fang, Cyrus Samii, Chris Schuck, Michael Schwam-Baird, and two anonymous reviewers are greatly appreciated.

*Address correspondence to Peter Aronow; E-mail: peter.aronow@yale.edu.

experiments. We demonstrate the effectiveness of our combined estimator and placebo tests with an original survey experiment.

KEY WORDS: Causal inference; Design-based inference; Political science; Randomized experiments; Sampling theory.

1. INTRODUCTION

The prevalence of sensitive attitudes and behaviors is difficult to estimate using standard survey techniques due to the tendency of respondents to withhold information in such settings. In recent years, the list experiment has grown in popularity as a method for eliciting truthful responses to sensitive questions. Introduced as the “item count technique” by Miller (1984), the procedure has been used to study racial prejudice (Kuklinski, Cobb, and Gilens 1997; Sniderman and Carmines 1997; Redlawsk, Tolbert, and Franko 2010), drug use (Biemer, Jordan, Hubbard, and Wright 2005; Coutts and Jann 2011), risky sexual activity (LaBrie and Earleywine 2000; Walsh and Braithwaite 2008), vote buying (Gonzalez-Ocantos, Kiewiet de Jonge, Meléndez, Osorio, and Nickerson 2012), and support for military occupation by foreign forces (Blair, Imai, and Lyall 2013). The standard list experiment proceeds by randomly partitioning respondents into control and treatment groups. Subjects in the control group receive a list of J non-sensitive items and report how many of the items apply to them. Subjects in the treatment group receive a list of $J + 1$ items comprised of the same J non-sensitive items plus one sensitive item. The list experiment estimate of the prevalence of the sensitive behavior is the difference-in-means between the treatment and control groups. The list experiment gives respondents cover to admit to engaging in the sensitive behavior—so long as the respondent reports between 1 and J items, the researcher cannot be certain whether an individual respondent engages in the sensitive behavior, but aggregate prevalence can be estimated.

List experiments may be useful because prevalence estimates based on direct questions are biased when some subjects tell the truth and others withhold information. In particular, the researcher cannot distinguish a respondent who does not engage in the sensitive behavior from one who does but is withholding: both types answer “No” to the direct question. Nevertheless, direct questions provide an important source of information when subjects *admit* to engaging in a sensitive behavior. Direct questions are biased but yield precise estimates of prevalence. Under certain assumptions, list experiments provide unbiased estimates of prevalence, but these estimates can be quite variable. The method we detail below allows researchers to reap the benefits of both direct questions and list experiments: increased precision and decreased bias. The central intuition of our approach is that, given a Monotonicity assumption (no false confessions), the true prevalence is a weighted average of two subject

types: those who admit to the sensitive behavior and those who withhold; we estimate the former with direct questions and the latter with list experiments.

A popular design for the list experiment is to randomly split the sample into three groups: those receiving the control list and no direct question, those receiving the treatment list and no direct question, and those receiving a direct question but no list at all (Brueckner, Morning, and Nelson 2005; Holbrook and Krosnick 2010; Heerwig and McCabe 2009). This design is often used so that direct and list experiment estimates can be compared within the same population. A variant of this design asks only subjects in the control group the direct question (Gilens, Sniderman, and Kuklinski 1998; Ahart and Sackett 2004). Our estimator requires that both treatment and control subjects receive a direct question. Examples of this more extensive measurement approach include Droitcour, Caspar, Hubbard, Parsley, Visscher, et al. (1991) and Gonzalez-Ocantos et al. (2012). Echoing similar design advice given in Kramon and Weghorst (2012) and Blair and Imai (2012), we advocate asking the direct question whenever feasible and investigating possible ordering effects.

When respondents are asked both direct and list questions, researchers can also test core assumptions underlying the list experiment: No Liars, No Design Effects, and ignorable treatment assignment (Imai 2011). The No Liars assumption requires that those who engage in the sensitive behavior do in fact include the sensitive item when reporting the number of list items that apply. The No Design Effects assumption requires that subjects' responses to the non-sensitive items on the list are unaffected by the presence or absence of the additional sensitive item. The treatment ignorability assumption requires that assignment be independent of both list experiment and direct question potential outcomes. These tests complement the one proposed by Blair and Imai (2012), which assesses whether any identified proportions of respondent types are negative, which would imply a contradiction between the model and the observed data.

We propose two tests. The logic of the first test, which is formalized below, is as follows: under the core list experiment assumptions and a Monotonicity assumption, the treatment versus control difference-in-means is in expectation equal to 1 among those who answer "Yes" to the direct question. Failing to reject the null hypothesis that the true difference-in-means for this subset is equal to 1 is equivalent to failing to reject the null hypothesis that the assumptions hold. We also propose a test of a variant of the ignorable treatment assignment assumption by assessing the dependence between responses to the direct question and the experimental treatment. While not conclusively demonstrating that the assumptions hold, these test results may give researchers more confidence that their survey instruments are providing reliable prevalence estimates.

Previous methodological work on list experiments has largely focused on two goals: decreasing the variance of list experiment estimates and modeling prevalence in a multivariate setting. Droitcour et al. (1991) propose the "Double List Experiment" design in which the prevalence of the same sensitive item is

investigated by two list experiments conducted with the same subjects, thereby reducing sampling variability. Holbrook and Krosnick (2010) use multivariate regression with treatment-by-covariate interaction terms to explore prevalence heterogeneity. Glynn (2013) suggests constructing the non-sensitive items so that they are negatively correlated with one another, a design feature that simultaneously reduces baseline variability and avoids ceiling effects. Corstange (2009) modifies the standard list experiment design by asking the control group each of the non-sensitive items directly, so that responses to the non-sensitive items can be modeled and more precise estimates of the sensitive items can be calculated. Imai (2011) proposes a nonlinear least squares estimator and a maximum likelihood estimator to model responses with covariate data. Blair and Imai (2012) offer a detailed review of these techniques.

Our contribution to the list experiment literature is to show the ease with which the additional information yielded by direct questioning can be incorporated into existing techniques. We demonstrate our proposed estimator and placebo tests on data from an original survey experiment conducted on Amazon's Mechanical Turk platform. We conclude with suggestions for the design and analysis of list experiments in scenarios where it is ethically feasible to ask direct questions as well.

2. SETTING AND IDENTIFICATION

Suppose we have a random sample of n subjects independently drawn from a large population. Let $X_i = 1$ if subject i engages in a sensitive behavior and $X_i = 0$ otherwise. We attempt to measure the behavior X_i using two methods: direct questioning and list experimentation. Our goal is to identify the prevalence of the sensitive behavior in the population, $\mu = \Pr[X_i = 1]$. Let Y_i be the report of subject i to the direct question. We assume that, under direct questioning, subjects may lie and claim that they do not engage in the behavior but will not lie and falsely claim that they do engage in the behavior.

Assumption 1 (Monotonicity). *There exist three latent classes of respondents under direct questioning: those who do not engage in the behavior and report truthfully ($X_i = 0, Y_i = 0$), subjects who engage and report truthfully ($X_i = 1, Y_i = 1$), and subjects who engage but report that they do not, i.e., withhold ($X_i = 1, Y_i = 0$).*

Let $p = \Pr[Y_i = 1 \mid X_i = 1]$ be the probability of a subject reporting truthfully to the direct question, given that he or she engages in the sensitive behavior. The response of subject i is

$$Y_i = \begin{cases} 0 & \text{with probability } 1 - \mu + \mu(1 - p) \\ 1 & \text{with probability } \mu p. \end{cases}$$

The response $Y_i = 0$ can be seen as a mixture of truthful negative reports and withholding. The probability that subject i engages in the behavior, given a negative response, is therefore

$$\Pr[X_i = 1 \mid Y_i = 0] = \frac{(1-p)\mu}{1-\mu p}.$$

While direct questioning is sufficient to reveal $\Pr[Y_i = 1] = \mu p$, it is not sufficient to identify $\Pr[Y_i = 0 \mid X_i = 1] = 1 - p$.

In contrast, the list experiment provides sufficient information to identify μ . Suppose we have a treatment $Z_i \in \{0, 1\}$. In the list experiment, treated subjects ($Z_i = 1$) receive a number of control questions and an additional question about the sensitive behavior. We denote the number of items that the subject states are applicable with V_i .

Assumption 2 (No Liars and No Design Effects). *Reframing Imai's (2011) formulations, we observe $V_i = W_i + X_i Z_i$, where W_i is the baseline outcome (under control) for subject i for the list experiment.*

We further require that the treatment assignment be independent of the actual behavior, direct question, and baseline response. This is a stricter variant of Imai's (2011) ignorability assumption.

Assumption 3 (Treatment Independence). $(W_i, X_i, Y_i) \perp\!\!\!\perp Z_i$.

Assumption 3 would be violated if (i) we did not have random assignment of the treatment or (ii) there are additional design effects; e.g., the treatment assignment affects the response to the direct question. Given random assignment, only the latter is a concern.

To ensure that all target quantities are well defined (and, later, to facilitate inference), we impose the mild regularity condition that all population variances are positive.

Assumption 4 (Non-Degenerate Distributions). $\text{Var}[V_i \mid Z_i = z, Y_i = y] > 0$, for $z, y \in \{0, 1\}$, $\text{Var}[Z_i] > 0$ and $\text{Var}[Y_i] > 0$.

We now turn to our primary identification result.

Lemma 1. *Given Assumptions 1–4 (Monotonicity, No Liars, No Design Effects, Treatment Independence, and Non-Degenerate Distributions), the prevalence may be represented as*

$$\mu = E[Y_i] + E[1 - Y_i](E[V_i \mid Z_i = 1, Y_i = 0] - E[V_i \mid Z_i = 0, Y_i = 0]). \quad (1)$$

Proof. By Assumptions 2 and 3,

$$E[V_i | Z_i = 1] - E[V_i | Z_i = 0] = W_i + \mu - W_i = \mu. \quad (2)$$

Then, expanding the left-hand side of (2) by marginalizing over Y_i , we represent the prevalence of the sensitive behavior as

$$\begin{aligned} \mu &= E[V_i | Y_i = 0, Z_i = 1]E[1 - Y_i] + E[V_i | Y_i = 1, Z_i = 1]E[Y_i] \\ &\quad - (E[V_i | Y_i = 0, Z_i = 0]E[1 - Y_i] + E[V_i | Y_i = 1, Z_i = 0]E[Y_i]) \\ &= (E[V_i | Y_i = 1, Z_i = 1] - E[V_i | Y_i = 1, Z_i = 0])E[Y_i] \\ &\quad + (E[V_i | Y_i = 0, Z_i = 1] - E[V_i | Y_i = 0, Z_i = 0])E[1 - Y_i] \\ &= E[X_i | Y_i = 1]E[Y_i] + (E[V_i | Y_i = 0, Z_i = 1] - E[V_i | Y_i = 0, Z_i = 0])E[1 - Y_i]. \end{aligned}$$

The result follows since $E[X_i | Y_i = 1] = 1$ by Assumption 2.

Note that if Assumptions 2–4 hold, but Assumption 1 (Monotonicity) does not hold, then $E[Y_i] + E[1 - Y_i](E[V_i | Z_i = 1, Y_i = 0] - E[V_i | Z_i = 0, Y_i = 0]) > \mu$, as then $E[X_i | Y_i = 1] < 1$.

3. ESTIMATION, INFERENCE, AND EFFICIENCY

In this section, we propose a simple nonparametric estimator of μ based on (1) and provide a basis for inference using Wald-type confidence intervals under a normal approximation. We also prove that our estimator is asymptotically more efficient than the standard difference-in-means estimator for the list experiment alone.

Define the sample means:

$$\begin{aligned} \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i \\ \bar{V}_{1,0} &= \frac{\sum_{i=1}^n V_i Z_i (1 - Y_i)}{\sum_{i=1}^n Z_i (1 - Y_i)} \quad \text{and} \quad \bar{V}_{0,0} = \frac{\sum_{i=1}^n V_i (1 - Z_i) (1 - Y_i)}{\sum_{i=1}^n (1 - Z_i) (1 - Y_i)}. \end{aligned}$$

Define an estimator of μ based on (1),

$$\hat{\mu} = \bar{Y} + (1 - \bar{Y})(\bar{V}_{1,0} - \bar{V}_{0,0}).$$

A preliminary lemma will assist us in deriving the asymptotic variance of this estimator.

Lemma 2. $\bar{V}_{1,0}$ and $\bar{V}_{0,0}$ are uncorrelated with $1 - \bar{Y}$.

A proof is given in Appendix B.

We can derive results on the sampling variance of $\hat{\mu}$. Let $\gamma = \Pr(Z_i = 1)$ be the probability of receiving the treatment question in the list experiment.

Proposition 1. *Given Assumption 4 (Non-Degenerate Distributions), the asymptotic variance of $\hat{\mu}$ is characterized by*

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} n\text{Var}[\hat{\mu}] &= \frac{(1 - \mu)^2}{1 - \mu p} \mu p + (1 - \mu p) \\ &\times \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 0]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 0]}{1 - \gamma} \right]. \end{aligned} \quad (3)$$

The proof is given in Appendix B. Under Assumptions 1–4, $\hat{\mu}$ is root- n consistent and asymptotically normal, with a consistent estimator of the variance obtained by substituting sample analogues (i.e., sample means and sample variances) for population quantities. Namely, let

$$\begin{aligned} n\widehat{\text{Var}}[\hat{\mu}] &= \frac{(1 - \hat{\mu})^2}{1 - \bar{Y}} \bar{Y} + (1 - \bar{Y}) \\ &\times \left[\frac{\hat{\sigma}^2(V_i | Z_i = 1, Y_i = 0)}{\hat{\gamma}} + \frac{\hat{\sigma}^2(V_i | Z_i = 0, Y_i = 0)}{1 - \hat{\gamma}} \right], \end{aligned}$$

where $\hat{\sigma}^2(\cdot)$ denotes the sample variance and $\hat{\gamma} = \sum_{i=1}^n Z_i/n$. These properties are sufficient for construction of Wald-type confidence intervals using $\widehat{\text{Var}}[\hat{\mu}]$.

Corollary 1. *If Assumptions 1–4 hold, then confidence intervals constructed as $\hat{\mu} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{\mu}]}$ will have μ $100(1 - \alpha)\%$ coverage for μ for large n .*

A proof follows directly from asymptotic normality and Slutsky’s theorem.

An estimator based upon (1) will have efficiency gains relative to standard difference-in-means-based estimators. Consider the standard difference-in-means-based estimator for the list experiment,

$$\hat{\mu}_S = \bar{V}_1 - \bar{V}_0,$$

where

$$\bar{V}_1 = \frac{\sum_{i=1}^n V_i Z_i}{\sum_{i=1}^n Z_i} \quad \text{and} \quad \bar{V}_0 = \frac{\sum_{i=1}^n V_i (1 - Z_i)}{\sum_{i=1}^n (1 - Z_i)}.$$

We now show that the combined estimator $\hat{\mu}$ is asymptotically more precise than $\hat{\mu}_S$.

Proposition 2. *Under Assumptions 1–4 (Monotonicity, No Liars, No Design Effects, Treatment Independence, and Non-Degenerate Distributions),*

$$\text{plim}_{n \rightarrow \infty} n\text{Var}[\hat{\mu}] < \text{plim}_{n \rightarrow \infty} n\text{Var}[\hat{\mu}_S].$$

The proof is given in Appendix B.

4. PLACEBO TESTS

In this section, we derive two placebo tests to assess the validity of the identifying assumptions.

4.1. Placebo Test I

It is possible to jointly test the Monotonicity, No Liars, No Design Effects, and Treatment Independence assumptions. Under these assumptions, for all t , $\Pr[V_i = t | Z_i = 0, Y_i = 1] = \Pr[V_i = (t + 1) | Z_i = 1, Y_i = 1]$, thus tests of distributional equality are appropriate. Any valid test of distributional equality between V_i (under $Z_i = 0, Y_i = 1$) and $V_i + 1$ (under $Z_i = 1, Y_i = 1$) will permit rejection of the null.

However, since distributional equality implies that $E[V_i | Z_i = 1, Y_i = 1] - E[V_i | Z_i = 0, Y_i = 1] = 1$, a simple test is available. Define $\beta = E[V_i | Z_i = 1, Y_i = 1] - E[V_i | Z_i = 0, Y_i = 1]$. Consider estimators

$$\hat{\beta} = \bar{V}_{1,1} - \bar{V}_{0,1}$$

and

$$\widehat{\text{Var}}[\hat{\beta}] = \frac{\hat{\sigma}^2(V_i | Z_i = 1, Y_i = 1)}{\sum_{i=1}^n Z_i Y_i} + \frac{\hat{\sigma}^2(V_i | Z_i = 0, Y_i = 1)}{\sum_{i=1}^n (1 - Z_i) Y_i}.$$

Proposition 3. *Under the null hypothesis that Assumptions 1–4 (Monotonicity, No Liars, No Design Effects, and Treatment Independence) hold, $\beta = 1$. For large n , if Assumption 4 (Non-Degenerate Distributions) holds, then a two-sided p -value is given by*

$$2\Phi\left(\frac{-|\hat{\beta} - 1|}{\sqrt{\widehat{\text{Var}}[\hat{\beta}]}}\right),$$

where $\Phi(\cdot)$ is the normal CDF.

A proof for Proposition 3 follows directly from calculations analogous to those for Proposition 1.

We also explore the power of Placebo Test I using a series of Monte Carlo simulations. We vary a number of factors, including the number of subjects answering “Yes” to the direct question, the proportions of false confessors, liars, and the design affected, and the variance of responses to the control list. The placebo test does not always have high power. For example, if 20 percent of 200 subjects responding “Yes” to the direct question are false confessors, the placebo test only has about 30 percent power. But when 20 percent of 800 subjects answering “Yes” are falsely confessing, the test has approximately

80 percent power. In general, the power of the placebo test depends on both the number of subjects answering “Yes” and the proportion of subjects violating the assumptions. These results are presented in Appendix C (please see the supplementary data online).

4.2 Placebo Test II

We can probe the validity of the Treatment Independence assumption with a second placebo test. Treatment Independence is violated if the answer to the direct question is systematically related to treatment assignment (i.e., $Y_i \not\perp Z_i$). Define $\delta = E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]$. Consider the estimators

$$\hat{\delta} = \frac{\sum_{i=1}^n Y_i Z_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n Y_i (1 - Z_i)}{\sum_{i=1}^n (1 - Z_i)}$$

and

$$\widehat{\text{Var}}[\hat{\delta}] = \frac{\hat{\sigma}^2(Y_i | Z_i = 1)}{\sum_{i=1}^n Z_i} + \frac{\hat{\sigma}^2(Y_i | Z_i = 0)}{\sum_{i=1}^n (1 - Z_i)}.$$

Proposition 4. *Under the null hypothesis that Assumption 3 holds, $\delta = 0$. For large n , if Assumption 4 (Non-Degenerate Distributions) holds, then a two-sided p -value is given by*

$$2\Phi\left(\frac{-|\hat{\delta}|}{\sqrt{\widehat{\text{Var}}[\hat{\delta}]}}\right).$$

A proof for Proposition 4 again follows directly from calculations analogous to those for Proposition 1. When the treatment is randomly assigned, Placebo Test II is simply a test of whether Z_i has a causal effect on Y_i . When Z_i is randomly assigned and the list experiment treatment is presented after the direct question, Assumption 3 holds by design.

5. APPLICATION

We tested the properties of our estimator with a pair of studies carried out on Amazon’s Mechanical Turk service, an Internet platform where subjects perform tasks such as the completion of surveys. Our main purpose was to assess the properties of our estimator by investigating an array of different behaviors, some of which may be considered socially sensitive. The relative anonymity of Internet surveys provides a favorable environment for list experiments precisely because we expect subjects to withhold less often than they might in face-to-face or telephone settings.

Table 1. Number of Subjects in Each Treatment Condition

		Study A or B		List 1		List 2		List 3		List 4		List 5	
		A	B	T	C	T	C	T	C	T	C	T	C
Study A or B	A	500	0	245	255	238	262	275	225	243	257	232	268
	B	0	514	272	242	259	255	253	261	271	243	260	254
List 1	T			517	0	251	266	273	244	269	248	273	244
	C			0	497	246	251	255	242	245	252	219	278
List 2	T					497	0	253	244	246	251	240	257
	C					0	517	275	242	268	249	252	265
List 3	T							528	0	269	259	263	265
	C							0	486	245	241	229	257
List 4	T									514	0	256	258
	C									0	500	236	264
List 5	T											492	0
	C											0	522

5.1 Experimental Design

We conducted five list experiments that paralleled five direct questions. The exact wording of the list experiments and direct questions is given in Appendix A. In three list experiments, we chose topics that are not socially sensitive: preferences over alternative energy sources, neighborhood characteristics, and news organizations. Two of the five list experiments dealt with racial and religious prejudice, topics where we would expect some withholding of anti-Hispanic and anti-Muslim sentiment.

We recruited a convenience sample of 1,023 subjects from Mechanical Turk. We offered subjects \$1.00 to complete our survey, which is equivalent to a \$15.45 hourly rate—a comparatively high wage by the standards of Mechanical Turk (Berinsky, Huber, and Lenz 2012). In order to defend against the potential for subjects to supply answers without reading or considering our questions, we included an “attention question” that required subjects to select a particular response in order to continue with the survey. Two subjects failed this quality check, and we exclude them from the main analysis. An additional seven subjects failed to respond to one or more of our questions, so we exclude them from the main analysis as well. The resulting sample size is $n = 1,014$.

Subjects were first assigned at random to either Study A or Study B. In Study A, direct questions were posed before the list questions, whereas in Study B, list questions were asked first. Subjects in both studies were then assigned to either the treatment or control conditions of each of the five list experiments. Table 1 displays the number of subjects in each treatment condition for each study, as well as every pairwise crossing of conditions.

Table 2. Covariate Balance: List Experiment 1

	Study A		Study B	
	Treat	Control	Treat	Control
18–24	24.49	25.49	23.53	32.64
25–34	40.82	43.92	41.18	40.91
35–44	20.00	15.29	17.28	14.88
45–54	8.16	10.98	10.66	6.20
55–64	4.90	3.92	5.88	3.72
65 or over	1.63	0.39	1.47	1.65
Female	46.12	47.45	46.32	42.98
Male	53.88	52.16	53.31	57.02
Prefer not to say—Gender	0.00	0.39	0.37	0.00
Liberal	50.61	42.35	50.37	51.65
Moderate	28.16	34.51	26.84	28.51
Conservative	18.37	19.22	20.59	14.46
Haven't thought much about this	2.86	3.92	2.21	5.37
Less than high school	0.41	0.78	0.74	1.24
High school/GED	11.02	11.76	10.66	7.44
Some college	42.45	42.75	40.44	40.91
4-Year college degree	33.47	33.73	35.29	36.36
Graduate school	12.65	10.98	12.87	14.05
White, non-Hispanic	79.18	77.65	79.78	80.17
African American	7.76	10.98	6.62	4.55
Asian/Pacific Islander	5.71	4.71	5.15	9.92
Hispanic	4.90	3.92	6.25	4.13
Native American	0.82	0.78	1.10	0.83
Other	1.22	1.18	1.10	0.41
Prefer not to say—Race	0.41	0.78	0.00	0.00
<i>n</i>	245	255	272	242

All randomizations used Bernoulli random assignment with equal probability 0.5. Consistent with our randomization procedure, each cell in the table (with the exception of the diagonal) contains approximately one-quarter of the subjects.

Before being randomized into treatment groups, subjects answered a series of background demographic questions. Table 2 shows balance statistics across age, gender, political ideology, education, and race across the treatment and control groups for the first list experiment in both studies. Our subject pool is more likely to be white, male, liberal, well educated, and young than the general population. This pattern is consistent with the demographic description of Mechanical Turk survey respondents given by [Mason and Suri \(2012\)](#).

Table 3. Study A (Directs First): Three Estimates of Prevalence

	Direct		Standard list		Combined list		% Reduction in sampling variance
	\bar{Y}	SE	$\hat{\mu}_s$	SE	$\hat{\mu}$	SE	
Nuclear power	0.656	0.021	0.748	0.084	0.666	0.048	67.1
Public transportation	0.538	0.022	0.513	0.072	0.627	0.049	54.0
Spanish speaking	0.102	0.014	0.035	0.079	0.042	0.074	14.0
Muslim teachers	0.110	0.014	0.166	0.081	0.187	0.074	15.3
CNN	0.444	0.022	0.338	0.105	0.533	0.070	54.9

NOTE. $n = 500$ for all estimates.

5.2 Study A (Direct Questions First)

In Study A, subjects were presented with the five direct questions before receiving the five list experiments. Table 3 presents three estimates of the prevalence in our subject pool. The first is a naive estimate computed by taking the average response to the direct question, \bar{Y} (Direct). The remaining two estimates are $\hat{\mu}_s$ (Standard List) and $\hat{\mu}$ (Combined List). For example, the direct question estimate \bar{Y} of the percentage agreeing that Muslims should not be allowed to teach in public schools¹ is 11 percent, the list experiment estimate $\hat{\mu}_s$ is 17 percent, and the combined estimate $\hat{\mu}$ is 19 percent. Of particular note are the standard errors associated with the standard list experiment as compared with those associated with the combined estimate: the reductions in estimated sampling variance are dramatic, ranging from 14 to 67 percent. As expected, reductions tend to be larger when a larger number of subjects respond “Yes” to the direct question. Figure 1 presents these results graphically: the estimates generally agree (providing confidence that the list experiments and the direct questions are measuring the same quantities), and the 95 percent confidence intervals around the combined estimate are always tighter than those around the standard estimate.

Table 4 presents the results of Placebo Test I. If these assumptions hold, the standard list experiment difference-in-means estimator will recover estimates that are in expectation equal to one among the subsample that answers “Yes” to the direct question. In two cases, we reject the joint null hypothesis of Monotonicity, No Liars, and No Design Effects: Public Transportation ($p = 0.02$) and CNN ($p = 0.03$). We speculate that some subjects may have felt

1. The pattern for the other socially sensitive topic, Spanish speaking, is reversed: the direct question estimate is greater than both list experimental estimates. Our replication study (described below) found the opposite pattern, suggesting that this apparent contrast is due to sampling variability.

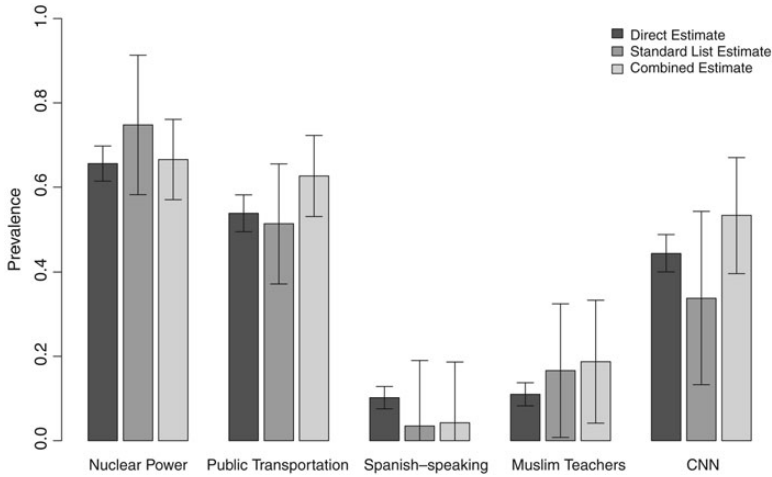


Figure 1. Study A (Directs First): Three Estimates of Prevalence.

Table 4. Study A (Directs First): Placebo Test I

	$\hat{\beta}$	SE	<i>p</i> -value	<i>n</i>
Nuclear power	1.054	0.095	0.568	328
Public transportation	0.790	0.091	0.021	269
Spanish speaking	0.848	0.279	0.585	51
Muslim teachers	1.008	0.237	0.973	55
CNN	0.696	0.143	0.034	222

that claiming to watch CNN was socially desirable, thereby violating Monotonicity.

Since we employed random assignment and the experimental treatment comes after the administration of the direct question, we expect to pass Placebo Test II, which seeks to verify that the treatment does not affect direct question responses. Indeed, as shown in table 5, the Placebo Test II results show no significant differences in mean responses to the direct questions by the list experimental treatment assignments.

5.3 Study B (List Experiments First)

Study B reverses the order of the direct questions and list experiments: subjects participated in all five list experiments before answering the direct questions. This design choice risks priming subjects in the treatment group in ways that might alter their responses to subsequent direct questions. For example, treated

Table 5. Study A (Directs First): Placebo Test II

	$\hat{\beta}$	SE	<i>p</i> -value	<i>n</i>
Nuclear power	0.066	0.042	0.118	500
Public transportation	-0.000	0.045	0.994	500
Spanish speaking	0.016	0.027	0.560	500
Muslim teachers	-0.030	0.028	0.285	500
CNN	-0.056	0.045	0.206	500

Table 6. Study B (Lists First): Three Estimates of Prevalence

	Direct		Standard list		Combined list		% Reduction in sampling variance
	\bar{Y}	SE	$\hat{\mu}_s$	SE	$\hat{\mu}$	SE	
Nuclear power	0.603	0.022	0.499	0.089	0.624	0.052	66.0
Public transportation	0.539	0.022	0.578	0.073	0.608	0.051	51.9
Spanish speaking	0.113	0.014	0.144	0.083	0.149	0.076	16.4
Muslim teachers	0.103	0.013	0.107	0.083	0.123	0.074	20.1
CNN	0.496	0.022	0.645	0.101	0.587	0.064	60.2

NOTE. *n* = 514 for all estimates.

subjects may be prone to misreport if subjects suspect that a particular topic is being given special scrutiny.

All three prevalence estimates for Study B are presented in table 6 and figure 2. The direct question estimates are very similar between Studies A and B—none of the differences between the estimates is significant at the 0.05 level. The standard list experiment estimates differ between Studies A and B, suggesting a question order effect. The combined estimator produces tighter estimates in Study B as well, with estimated sampling variability reductions in a very similar range. Appendix E (please see the supplementary data online) presents formal tests of the differences in estimates across the studies.

The results of Placebo Test I for Study B are presented in table 7. Among the subgroup of respondents who answer “Yes” to the direct question, the list experiment difference-in-means estimate β should be equal to 1, under Assumptions 1–4 (Monotonicity, No Liars, No Design Effects, and Treatment Independence). None of the values of β are statistically significantly different from 1 using the placebo test, and a joint test via Fisher’s method is insignificant as well.

As described in section 4.2, the combined estimator relies in part on the assumption that a subject’s response to the direct question is unaffected by the

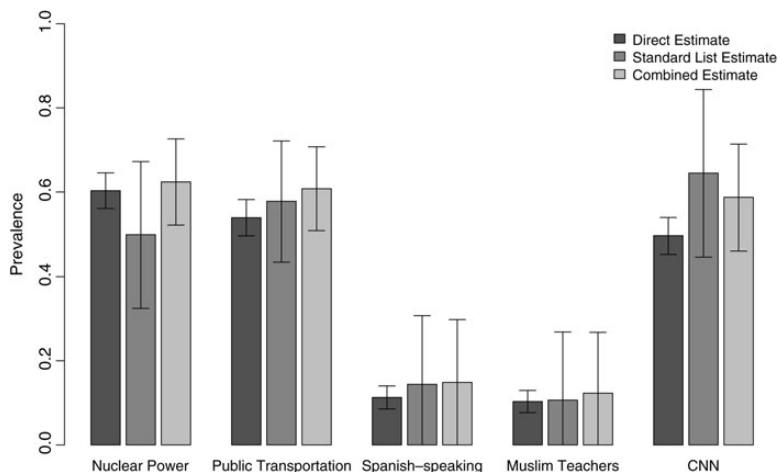


Figure 2. Study B (Lists First): Three Estimates of Prevalence.

Table 7. Study B (Lists First): Placebo Test I

	$\hat{\beta}$	SE	<i>p</i> -value	<i>n</i>
Nuclear power	0.881	0.113	0.294	310
Public transportation	0.913	0.091	0.339	277
Spanish speaking	0.767	0.229	0.309	58
Muslim teachers	0.700	0.285	0.293	53
CNN	0.847	0.135	0.256	255

list experimental treatment assignment. Violations of this assumption are directly testable using Placebo Test II. In Study B, subjects were exposed to either a treatment or a control list before answering the direction question. Table 8 below presents the effect the treatment lists may have had on answers to the direct questions. In two of the five cases, direct questions were significantly affected by the treatment list: treated subjects were 8.6 percentage points less likely to declare their support for nuclear power and were 13.2 percentage points more likely to report watching CNN. These findings indicate that the Treatment Independence assumption is most likely violated for these questions, rendering the Study B combined list estimates for these two questions unreliable.

5.4 Replication Study

We conducted a replication study following the identical design with 506 new Mechanical Turk subjects in a replication of Study A and 506 in a replication of

Table 8. Study B (Lists First): Placebo Test II

	$\hat{\beta}$	SE	<i>p</i> -value	<i>n</i>
Nuclear power	-0.086	0.043	0.045	514
Public transportation	0.034	0.044	0.434	514
Spanish speaking	0.027	0.028	0.337	514
Muslim teachers	0.016	0.027	0.549	514
CNN	0.132	0.044	0.003	514

Study B. Full results of this replication study are presented in Appendix F (please see the supplementary data online), but the findings are strikingly similar to the first investigation. The list experimental estimates vary somewhat between the original experiments and the replication, but none of these differences are significant ($p > 0.05$). One of the five placebo tests was significant in Study A, and three of the five were significant in Study B. Perhaps surprisingly, the effect of the treatment list on direct answers to the CNN question presented in table 8 was also replicated.

6. DISCUSSION

Social desirability effects may bias prevalence estimates of sensitive behaviors and opinions obtained using direct questioning, but that does not mean that direct questions are useless. Under an assumption of Monotonicity (subjects who do not engage in the sensitive behavior do not falsely confess), direct questions reveal reliable information about those who answer “Yes.” Among those who answer “No,” we cannot directly distinguish those who withhold from those who do not engage in the sensitive behavior—for these subjects, list experiments may provide a workaround. Combining these two techniques into a single estimator yields more precise estimates of prevalence, and employing direct and list questions in tandem also enables the researcher to test crucial identifying assumptions.

A few caveats are in order with respect to empirical applications. First, Monotonicity is not guaranteed to hold, especially when social desirability cuts in opposite directions for different subgroups. For example, moderates in liberal areas may feel pressure to support Muslim teachers, whereas moderates in conservative areas may feel pressure to oppose them. Second, list experiments are often employed when the safety of respondents would be compromised if they admitted to sensitive opinions or behaviors (e.g., Pashtun respondents admitting support for NATO forces; Blair et al. 2013). We do not take these concerns lightly, and in such cases would not recommend the use of our method. Third, the order in which direct questions and list experiments are asked appears to matter. Unfortunately, the empirical results of Placebo Test I

fail to provide clear guidance with respect to ordering: we reject the joint null hypothesis of Monotonicity, No Liars, and No Design Effects for two of the experiments in Study A but fail to reject it for any of the five experiments in Study B. Our replication study saw the opposite pattern: one rejection in Study A, and three rejections in Study B. Placebo Test II, on the other hand, suggests that, at least in our application, asking the direct question second induced a violation of the Treatment Independence assumption. In sum, we recommend randomizing the order in which the list experiment and the direct question are presented so that (a) question-order effects can be contained and (b) the relevant tests of the assumptions can be performed. Finally, the power of Placebo Test I varies with the prevalence rate, and is consequently less useful when the goal of the list experiment is to estimate the prevalence of a rare attitude or behavior—a common circumstance if one imagines that sensitive behaviors also tend to be low prevalence. Nevertheless, Placebo Test I detected many instances of violated assumptions (six of twenty opportunities), suggesting that caution is warranted when interpreting list experimental estimates of prevalence.

We have combined direct question estimates with the simplest of the various list experiment estimators: difference-in-means. This work could be extended straightforwardly to the multivariate settings discussed by [Corstange \(2009\)](#), [Holbrook and Krosnick \(2010\)](#), and [Imai \(2011\)](#). One such approach would involve regression estimation ([Särndal, Swensson, and Wretman 1992](#); [Lin 2013](#)) or post-stratification ([Holt and Smith 1979](#); [Miratrix, Sekhon, and Yu 2013](#)) for computing covariate adjusted means. Such an approach would improve asymptotic efficiency without any parametric assumptions, and a consistent variance estimator may be derived by substituting residuals from the regression fit. Finally, we note that other methods for eliciting truthful responses to sensitive questions, such as randomized response ([Warner 1965](#)) and endorsement experiments ([Bullock, Imai, and Shapiro 2011](#); [Lyall, Blair, and Imai 2013](#)), could also be combined with direct questioning to improve precision.

Supplementary Materials

Supplementary materials are available online at http://www.oxfordjournals.org/our_journals/jssam/.

Appendix A.1

Survey Text

Table A1. Survey Questions and Response Options

	List question (sensitive item in bold)	Direct question
Nuclear power (L1 & D1)	<p>Here is a list of four (five) alternative energy sources that some people think the United States should be developing. Please read all four (five) and select how many of them you think the United States should be developing. We do not want to know which ones, just how many.</p> <p>Wind</p> <p>Solar</p> <p>Nuclear power</p> <p>Natural gas</p> <p>Ethanol</p>	<p>Some people think nuclear power is an alternative energy source that the United States should be developing. How about you? Do you think nuclear power is an alternative energy source that the United States should be developing?</p> <p>Yes</p> <p>No</p>
Public transportation (L2 & D2)	<p>Here is a list of three (four) things some people consider important when choosing a neighborhood to live in. Please read all three (four) and select how many of them you consider important when choosing a neighborhood. We do not want to know which ones, just how many.</p> <p>Proximity to shops and restaurants</p> <p>Quality of local schools</p> <p>Neighbors who share my values</p> <p>Access to public transportation</p>	<p>Some people consider access to public transportation important when choosing a neighborhood to live in. How about you? Do you consider access to public transportation important when choosing a neighborhood to live in?</p> <p>Yes</p> <p>No</p>

Spanish speaking (L3 & D3)	<p>Here is a list of four (five) things that sometimes make people angry or upset. Please read all four (five) and select how many of them upset you. We do not want to know which ones, just how many.</p> <p>The federal government increasing the tax on gasoline Professional athletes earning large salaries</p> <p>A Spanish-speaking family moving in next door Requiring seat belts be used when driving Large corporations polluting the environment</p>	<p>Sometimes people are angry or upset when a Spanish-speaking family moves in next door. How about you? Would a Spanish-speaking family moving in next door upset you?</p> <p>Yes No</p>
Muslim teachers (L4 & D4)	<p>Here is a list of three (four) groups that some people think should be prohibited from teaching in public schools. We do not want to know which ones, just how many.</p> <p>18–21-year-olds Muslims People without a master’s degree in education People who earn a 2.0 GPA or lower</p>	<p>Sometimes people think that Muslims should be prohibited from teaching in public schools. How about you? Do you think that Muslims should be prohibited from teaching in public schools?</p> <p>Yes No</p>
CNN (L5 & D5)	<p>Here is a list of four (five) news organizations. Please read all four (five) and select how many you read or watch in the course of an ordinary month. We do not want to know which ones, just how many.</p> <p><i>The New York Times</i> CNN The Huffington Post Fox News Politico</p>	<p>In the course of an ordinary month, do you watch CNN?</p> <p>Yes No</p>

Appendix B.1

Proofs

Proof of Lemma 2

Proof. To show that $\bar{V}_{1,0}$ and $\bar{V}_{0,0}$ are uncorrelated with $1 - \bar{Y}$, we demonstrate that the expected values of these variables are invariant to conditioning on $1 - \bar{Y}$. Suppose that exactly k direct responses are zero, so $1 - \bar{Y} = k/n$. Then,

$$\begin{aligned} E\left[\bar{V}_{1,0} \mid 1 - \bar{Y} = \frac{k}{n}\right] &= E\left[\frac{\sum_{j=1}^n V_j Z_j (1 - Y_j)}{\sum_{j=1}^n Z_j (1 - Y_j)} \mid 1 - \bar{Y} = \frac{k}{n}\right] \\ &= E\left[\frac{\sum_{j=1}^k V_j Z_j}{\sum_{j=1}^k Z_j} \mid Y_j = 0 \text{ for } j = 1, \dots, k\right], \end{aligned} \quad (4)$$

where we have reordered the indices so that $Y_1 = \dots = Y_k = 1$. Then, applying the law of iterated expectation, we have

$$\begin{aligned} E\left[\bar{V}_{1,0} \mid 1 - \bar{Y} = \frac{k}{n}\right] &= E_Z \left[E\left[\frac{\sum_{j=1}^k V_j Z_j}{\sum_{j=1}^k Z_j} \mid Y_j = 0 \text{ for } j = 1, \dots, k, \sum_{j=1}^k Z_j = i\right] \right] \\ &= \sum_{i=1}^k \frac{\sum_{j=1}^i E[V_j \mid Z_j = 1, Y_j = 0]}{i} \Pr\left(\sum_{j=1}^k Z_j = i\right), \end{aligned} \quad (5)$$

where we have again reordered the indices so that $Z_1 = \dots = Z_i = 1$. Since $E[V_j \mid Z_j = 1, Y_j = 0]$ is the same for every $j = 1, \dots, i$,

$$\begin{aligned} E\left[\bar{V}_{1,0} \mid 1 - \bar{Y} = \frac{k}{n}\right] &= E[V_1 \mid Z_1 = 1, Y_1 = 0] \sum_{i=1}^k \Pr\left(\sum_{j=1}^k Z_j = i\right) \\ &= E[V_1 \mid Z_1 = 1, Y_1 = 0]. \end{aligned} \quad (6)$$

Then, since the last line does not depend on k , we conclude that

$$E\left[\bar{V}_{1,0} \mid 1 - \bar{Y} = \frac{k}{n}\right] = E\left[\bar{V}_{1,0} \mid 1 - \bar{Y} = \frac{k'}{n}\right]$$

for $k \neq k'$. It follows that the expectation of $\bar{V}_{1,0}$ is invariant to conditioning on $1 - \bar{Y}$, and so $E[\bar{V}_{1,0} \mid 1 - \bar{Y}] = E[\bar{V}_{1,0}]$.

A similar argument holds for $\bar{V}_{0,0}$:

$$\begin{aligned}
 E\left[\bar{V}_{0,0} \mid 1 - \bar{Y} = \frac{k}{n}\right] &= E\left[\frac{\sum_{j=1}^n V_j(1 - Z_j)(1 - Y_j)}{\sum_{j=1}^n (1 - Z_j)(1 - Y_j)} \mid 1 - \bar{Y} = \frac{k}{n}\right] \\
 &= E\left[\frac{\sum_{j=1}^k V_j(1 - Z_j)}{\sum_{j=1}^k (1 - Z_j)} \mid Y_j = 0 \text{ for } j = 1, \dots, k\right] \\
 &= E_Z\left[E\left[\frac{\sum_{j=1}^k V_j(1 - Z_j)}{\sum_{j=1}^k (1 - Z_j)} \mid Y_j = 0 \text{ for } j = 1, \dots, k, \sum_{j=1}^k (1 - Z_j) = i\right]\right] \\
 &= \sum_{i=1}^k \frac{\sum_{j=1}^i E[V_j \mid Z_j = 0, Y_j = 0]}{i} \Pr\left(\sum_{j=1}^k (1 - Z_j) = i\right) \\
 &= E[V_1 \mid Z_1 = 0, Y_1 = 0] \sum_{i=1}^k \Pr\left(\sum_{j=1}^k (1 - Z_j) = i\right) \\
 &= E[V_1 \mid Z_1 = 0, Y_1 = 0].
 \end{aligned}
 \tag{7}$$

Since the expectations of both $\bar{V}_{1,0}$ and $\bar{V}_{0,0}$ are unchanged by conditioning on $1 - \bar{Y}$, these variables are uncorrelated with $1 - \bar{Y}$, as claimed.

Proof of Proposition 1

Proof. The proof proceeds by working with linearized variances:

$$\begin{aligned}
 \text{Var}[\hat{\mu}] &= \text{Var}[1 + (1 - \bar{Y})(\bar{V}_{1,0} - \bar{V}_{0,0} - 1)] \\
 &= \text{Var}[(1 - \bar{Y})(\bar{V}_{1,0} - \bar{V}_{0,0} - 1)].
 \end{aligned}
 \tag{8}$$

By Lemma 2, $\bar{V}_{1,0}$ and $\bar{V}_{0,0}$ are uncorrelated, so the variance of the product in (8) decomposes as follows:

$$\begin{aligned}
 \text{Var}[\hat{\mu}] &= (E[1 - \bar{Y}])^2 \text{Var}[\bar{V}_{1,0} - \bar{V}_{0,0}] + \text{Var}[1 - \bar{Y}](E[\bar{V}_{1,0} - \bar{V}_{0,0} - 1])^2 \\
 &\quad + \text{Var}[1 - \bar{Y}]\text{Var}[\bar{V}_{1,0} - \bar{V}_{0,0}] \\
 &= (1 - \mu p)^2 \text{Var}[\bar{V}_{1,0} - \bar{V}_{0,0}] + \frac{\mu p(1 - \mu p)}{n} \left(\frac{1 - \mu}{1 - \mu p}\right)^2 \\
 &\quad + \frac{\mu p(1 - \mu p)}{n} \text{Var}[\bar{V}_{1,0} - \bar{V}_{0,0}] + O(n^{-2}) \\
 &= \frac{\mu p(1 - \mu)^2}{n(1 - \mu p)} + (1 - \mu p)^2 \left[\frac{\text{Var}[V_i \mid Z_i = 1, Y_i = 0]}{(1 - \mu p)n\gamma} + \frac{\text{Var}[V_i \mid Z_i = 0, Y_i = 0]}{(1 - \mu p)(n - n\gamma)}\right] + O(n^{-2}) \\
 &= \frac{\mu p(1 - \mu)^2}{n(1 - \mu p)} + (1 - \mu p) \left[\frac{\text{Var}[V_i \mid Z_i = 1, Y_i = 0]}{n\gamma} + \frac{\text{Var}[V_i \mid Z_i = 0, Y_i = 0]}{n - n\gamma}\right] + O(n^{-2}),
 \end{aligned}
 \tag{9}$$

where γ is the probability of receiving treatment, so $n\gamma = \sum_{i=1}^n Z_i$. Multiplying by n yields the desired result.

Proof of Proposition 2

Proof. We begin by expressing the asymptotic variance of $\hat{\mu}_S$:

$$\begin{aligned}
 \text{plim}_{n \rightarrow \infty} n \text{Var}[\hat{\mu}_S] &= (1 - \mu p) \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 0]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 0]}{1 - \gamma} \right] \\
 &\quad + \mu p \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 1]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 1]}{1 - \gamma} \right] \\
 &\quad + (1 - \mu p) \left[\frac{(\text{E}[V_i | Z_i = 1, Y_i = 0] - \text{E}[V_i | Z_i = 1])^2}{\gamma} \right] \\
 &\quad + (1 - \mu p) \left[\frac{(\text{E}[V_i | Z_i = 0, Y_i = 0] - \text{E}[V_i | Z_i = 0])^2}{1 - \gamma} \right] \\
 &\quad + \mu p \left[\frac{(\text{E}[V_i | Z_i = 1, Y_i = 1] - \text{E}[V_i | Z_i = 1])^2}{\gamma} \right] \\
 &\quad + \mu p \left[\frac{(\text{E}[V_i | Z_i = 0, Y_i = 1] - \text{E}[V_i | Z_i = 0])^2}{1 - \gamma} \right] \\
 &= (1 - \mu p) \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 0]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 0]}{1 - \gamma} \right] \\
 &\quad + \mu p \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 1]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 1]}{1 - \gamma} \right] \\
 &\quad + \mu p(1 - \mu p) \left[\frac{(\text{E}[V_i | Z_i = 1, Y_i = 0] - \text{E}[V_i | Z_i = 1, Y_i = 1])^2}{\gamma} \right] \\
 &\quad + \mu p(1 - \mu p) \left[\frac{(\text{E}[V_i | Z_i = 0, Y_i = 0] - \text{E}[V_i | Z_i = 0, Y_i = 1])^2}{1 - \gamma} \right].
 \end{aligned}$$

By Assumptions 1 and 2 (Monotonicity, No Liars, and No Design Effects), $\text{E}[V_i | Z_i = 1, Y_i = 1] = \text{E}[V_i | Z_i = 0, Y_i = 1] + 1$. Then,

$$\begin{aligned}
 \text{plim}_{n \rightarrow \infty} n \text{Var}[\hat{\mu}_S] &= (1 - \mu p) \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 0]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 0]}{1 - \gamma} \right] \\
 &\quad + \mu p \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 1]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 1]}{1 - \gamma} \right] \\
 &\quad + \mu p(1 - \mu p) \left[\frac{((\mu - 1)/(1 - \mu p) + \text{E}[V_i | Z_i = 0, Y_i = 0] - \text{E}[V_i | Z_i = 0, Y_i = 1])^2}{\gamma} \right] \\
 &\quad + \mu p(1 - \mu p) \left[\frac{(\text{E}[V_i | Z_i = 0, Y_i = 0] - \text{E}[V_i | Z_i = 0, Y_i = 1])^2}{1 - \gamma} \right].
 \end{aligned}$$

Applying the first-order condition, $\text{plim}_{n \rightarrow \infty} n \text{Var}[\hat{\mu}_S]$ is minimized when $\text{E}[V_i | Z_i = 0, Y_i = 0] - \text{E}[V_i | Z_i = 0, Y_i = 1] = [\gamma - 1][(\mu - 1)/(1 - \mu p)]$.

Substituting terms, it follows that

$$\begin{aligned}
 \text{plim}_{n \rightarrow \infty} n \text{Var}[\hat{\mu}_s] &\geq (1 - \mu p) \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 0]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 0]}{1 - \gamma} \right] \\
 &+ \mu p \left[\frac{\text{Var}[V_i | Z_i = 1, Y_i = 1]}{\gamma} + \frac{\text{Var}[V_i | Z_i = 0, Y_i = 1]}{1 - \gamma} \right] \\
 &+ \mu p (1 - \mu p) \left(\frac{1 - \mu}{1 - \mu p} \right)^2 \\
 &\geq \text{plim}_{n \rightarrow \infty} n \text{Var}[\hat{\mu}].
 \end{aligned}
 \tag{10}$$

Assumption 4 (Non-Degenerate Distributions) ensures that the inequality holds strictly.

References

- Ahart, A. M., and P. R. Sackett (2004), "A New Method of Examining Relationships between Individual Difference Measures and Sensitive Behavior Criteria: Evaluating the Unmatched Count Technique," *Organizational Research Methods*, 7(1), 101–114.
- Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012), "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, 20(3), 351–368.
- Biemer, P. P., B. K. Jordan, M. L. Hubbard, and D. Wright (2005), "A Test of the Item Count Methodology for Estimating Cocaine Use Prevalence," in *Evaluating and Improving Methods Used in the National Survey on Drug Use and Health*, eds. J. Kennet and J. Gfroerer, Department of Health and Human Services.
- Blair, G., and K. Imai (2012), "Statistical Analysis of List Experiments," *Political Analysis*, 20(1), 47–77.
- Blair, G., K. Imai, and J. Lyall. (2013), "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan," unpublished manuscript.
- Brueckner, H., A. Morning, and A. Nelson (2005), "The Expression of Biological Concepts of Race," unpublished manuscript.
- Bullock, W., K. Imai, and J. N. Shapiro (2011), "Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan," *Political Analysis*, 19(4), 363–384.
- Corstange, D. (2009), "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT," *Political Analysis*, 17(1), 45–63.
- Coutts, E., and B. Jann (2011), "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)," *Sociological Methods & Research*, 40(1), 169–193.
- Droitcour, J., R. A. Caspar, M. L. Hubbard, T. L. Parsley, W. Visscher, and T. M. Ezzati (1991), "The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application," in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, chapter 11, pp. 185–210, Hoboken, NJ: John Wiley & Sons.
- Gilens, M., P. M. Sniderman, and J. H. Kuklinski (1998), "Affirmative Action and the Politics of Realignment," *British Journal of Political Science*, 28(1), 159–183.
- Glynn, A. N. (2013), "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment," *Public Opinion Quarterly*, 77(S1), 159–172.

- Gonzalez-Ocantos, E., C. Kiewiet de Jonge, C. Meléndez, J. Osorio, and D. W. Nickerson (2012), "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua," *American Journal of Political Science*, 56(1), 202–217.
- Heerwig, J. A., and B. J. McCabe (2009), "Education and Social Desirability Bias: The Case of a Black Presidential Candidate," *Social Science Quarterly*, 90(3), 674–686.
- Holbrook, A. L., and J. A. Krosnick (2010), "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique," *Public Opinion Quarterly*, 74(1), 37–67.
- Holt, D., and T. M. F. Smith (1979), "Post Stratification," *Journal of the Royal Statistical Society. Series A*, 142(1), 33–46.
- Imai, K. (2011), "Multivariate Regression Analysis for the Item Count Technique," *Journal of the American Statistical Association*, 106(494), 407–416.
- Kramon, E., and K. R. Weghorst (2012), "Measuring Sensitive Attitudes in Developing Countries: Lessons from Implementing the List Experiment," *Newsletter of the APSA Experimental Section*, 3(2), 14–24.
- Kuklinski, J. H., M. D. Cobb, and M. Gilens (1997), "Racial Attitudes and the New South," *Journal of Politics*, 59(2), 323–349.
- LaBrie, J. W., and M. Earleywine (2000), "Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique," *Journal of Sex Research*, 37(4), 321–326.
- Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *Annals of Applied Statistics*, 7(1), 295–318.
- Lyall, J., G. Blair, and K. Imai (2013), "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan," *American Political Science Review*, 107(4), 679–705.
- Mason, W., and S. Suri (2012), "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods*, 44(1), 1–23.
- Miller, J. D. (1984), "A New Survey Technique for Studying Deviant Behavior," Ph.D. thesis, George Washington University.
- Miratrix, L. W., J. S. Sekhon, and B. Yu (2013), "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments," *Journal of the Royal Statistical Society, Series B (Methodology)*, 75(2), 369–396.
- Redlawsk, D. P., C. J. Tolbert, and W. Franko (2010), "Voters, Emotions, and Race in 2008: Obama as the First Black President," *Political Research Quarterly*, 63(4), 875–889.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling*, New York: Springer.
- Sniderman, P. M., and E. G. Carmines (1997), *Reaching Beyond Race*, Cambridge, MA: Harvard University Press.
- Walsh, J. A., and J. Braithwaite (2008), "Self-Reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-Count Technique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students," *Journal of Alcohol and Drug Education*, 52(2), 49–72.
- Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60(309), 63–69.