Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence

Online Supplementary Materials

Peter M. Aronow, Alexander Coppock, Forrest W. Crawford,
and Donald P. Green*

August 7, 2014

*Peter M. Aronow is Assistant Professor, Department of Political Science, Yale University, 77 Prospect Street, New Haven, CT 06520. Alexander Coppock is Doctoral Student, Department of Political Science, Columbia University, 420 W. 118th Street, New York, NY 10027. Forrest W. Crawford is Assistant Professor, Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520. Donald P. Green is Professor of Political Science at Columbia University, 420 W. 118th Street, New York, NY 10027. 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). Helpful comments from Xiaoxuan Cai, Albert Fang, Cyrus Samii, Chris Schuck, Michael Schwam-Baird, and two anonymous reviewers are greatly appreciated.

Appendix C: Placebo Test I Power Analysis

The placebo test proposed in Section 4 is a joint test of Assumptions 1-3 (Monotonicity, No Liars, No Design Effects, and Treatment Independence). In brief, the test considers whether the conventional list experimental estimate appears to be significantly different from 1.0 among the subset of subjects who answer "Yes" to the direct question. If this estimate is different from 1.0, it must either be because a) some of those who answer "Yes" are falsely confessing (thereby violating monotonicity) or b) the standard list experiment assumptions of No Liars and No Design Effects are not met. We vary the following five quantities: the number of subjects who answer "Yes" to the direct question (N_{Yes}, or $\sum_{i=1}^{N} Y_i$) the variability of responses to the non-sensitive list items, the proportion of N_{Yes} who falsely confess, the proportion of N_{Yes} who lie when given the treatment list, and the proportion of N_{Yes} whose responses to the non-sensitive list items change when given the treatment list.

We display the results of four power simulations in Figure A1 below. On the y-axis of each panel, we vary N_{Yes}. On the x-axis of the first three panels, we vary the proportion of those subjects whose response profile violates one of the assumptions: No False Confessions, No Liars, or No Design Effects, respectively. The proportion of units that violate the other two assumptions was fixed at 0. Control list responses were drawn from a binomial distribution with a success probability of 0.4 and four trials. For those units who do meet the assumptions of No False Confessions, No Liars, or No Design Effects, treatment list responses were set equal to the control list response plus one. The treatment list responses for Liars and False Confessors were set equal to their control list responses. The treatment list response for Design Affected subjects was generated as a "ceiling effect": the control list plus one, except for those with a "4" on the control list; those units' treatment list response remained equal to 4.[1] The final panel fixes the proportion of false confessors at 0.20 and changes the variability of responses to the non-sensitive list items. We parameterized the variability in responses to non-sensitive list items as the success probability of a binomial distribution with four trials. This variability is maximized when $p = 0.5$.

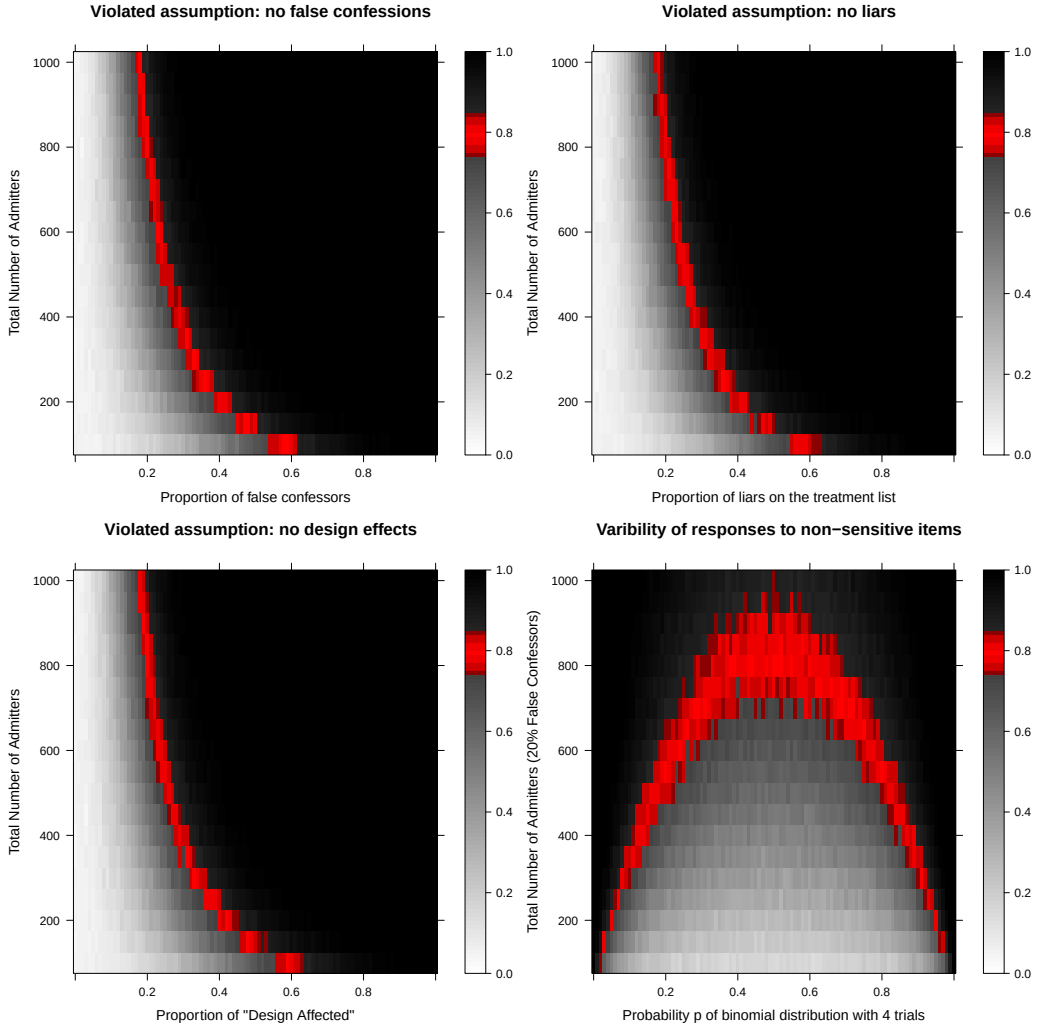
We varied N_{Yes} from 100 to 1000 in steps of 50, the proportion of False Confessors, Liars, and Design Affected units from 0 to 1 in steps of 0.01, and the success probability of the binomial distribution from 0 to 1 in steps of 0.01. We conducted 1,000 simulations of each combination. The shading reflects the proportion of simulations in which we were able to reject the null hypothesis that Assumptions 1 and 2 hold, with darker shades corresponding to higher power. For ease of interpretation, the shading around the conventional power target of 0.80 is shaded red.

The simulations show that for any level of violation, a larger sample size increases the power of the test. At any sample size, greater proportions of violators increase the power

[1] This is one of many possible design effects; the power of our test to detect any particular design effect depends on the manner in which it changes subjects' treatment list responses.

of the test. When approximately 20% of a sample of 800 confessors falsely confess, lie on the treatment list, or change their responses to the non-sensitive items on the treatment list, the placebo test achieves 80% power. The final panel shows that the power of the test is maximized when the variability in non-sensitive list items is minimized.

Figure A1: Power of Placebo Test I to Detect Violations of Assumptions 1-3

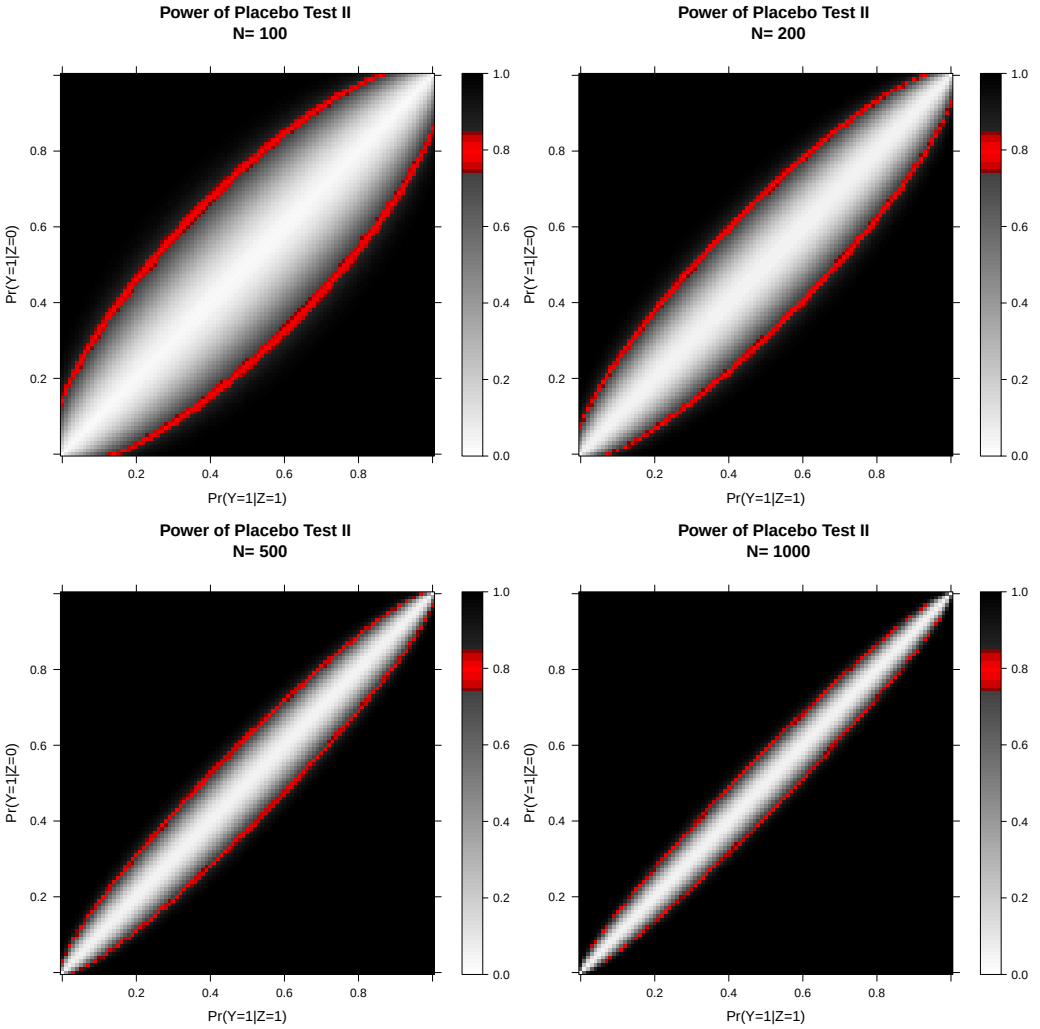


Appendix D: Placebo Test II Power Analysis

The power of Placebo Test II to detect violations of Treatment Independence is equivalent to the power of test of a difference in proportions. Following the implementation in R (see accompanying text: Dalgaard 2008, p. 159), the power of the test for N = 100, 200, 500, and 1000 is shown in Figure A2 below. In all panels, the proportion of "Yes" responses in the treatment group is given on the x-axis, and the proportion in the control group is

given on the y-axis. Darker coloring indicates higher power, with the band around the 0.80 power target shaded red.

Figure A2: Power of Placebo Test II to Detect Violations of Treatment Independence



Appendix E: Factorial Design

Subjects were randomly assigned to Study A (Direct Questions first) or Study B (List Experiments First). Table A1 below presents the differences in estimates obtained by the two studies. The order in which the direct and list questions does not appear to have had a significant impact on the Direct or Combined List estimates of prevalance, but does appear to have changed the Conventional List estimates in two cases. The list experimental estimate of the proportion supporting Nuclear power is 24.9 percentage points higher

when the direct questions are asked first, whereas the estimate of those watching CNN is 30.7 percentage points lower when the direct questions are asked first.

Table A1: Factorial Design: Study A - Study B

	Direct		Conventional List		Combined List	
	Difference	SE	Difference	SE	Difference	SE
Nuclear Power	0.053	0.030	0.249	0.122	0.042	0.071
Public Transportation	-0.001	0.031	-0.064	0.103	0.019	0.071
Spanish-speaking	-0.011	0.019	-0.109	0.115	-0.106	0.106
Muslim Teachers	0.007	0.019	0.060	0.116	0.064	0.105
CNN	-0.052	0.031	-0.307	0.146	-0.054	0.095

Appendix F: Replication Study

The replication study was conducted with a new pool of Mechanical Turk respondents ten months after the first study. The experimental procedure was identical in every respect to the design described in Section 5. 1019 users started the survey, but 7 either did not complete the survey or failed the attention question, leaving 1012 complete cases. 506 subjects participated in Study A and 506 participated in Study B.

The formats of the tables below follow those Section 5, facilitating comparisons. Of particular note are the results of the placebo tests, presented in Tables A3, A6, and A7. In the original study, two of five questions failed the placebo test in Study A and none of the question failed in Study B. The replication shows a different pattern: one of five fail in Study A, whereas four of five (at the 10% level or greater) fail in Study B. In the original study, both Nuclear Power and CNN failed the second placebo test, but in the replication, only CNN fails. We interpret this result to mean that the effect of the treatment list on direct answers to the CNN is not a fluke due to sampling variability but rather a robust causal effect on the order of 10 percentage points.

Table A2: Study A (Direct First): Three Estimates of Prevalence

	Direct		Standard List		Combined List		% Reduction in Sampling Variance
	$\overline{Y}$	SE	$\hat{\mu}_S$	SE	$\hat{\mu}$	SE	
Nuclear Power	0.646	0.021	0.575	0.086	0.647	0.049	67.326
Public Transportation	0.555	0.022	0.635	0.073	0.653	0.048	56.643
Spanish-speaking	0.061	0.011	0.118	0.078	0.104	0.073	11.798
Muslim Teachers	0.083	0.012	0.034	0.078	0.036	0.072	14.651
CNN	0.407	0.022	0.256	0.100	0.298	0.074	45.227

n = 506 for all estimates

Table A3: Study A (Direct First): Placebo Test I

	$\hat{\beta}$	SE	p-value	n
Nuclear Power	0.938	0.104	0.549	327.000
Public Transportation	0.949	0.089	0.566	281.000
Spanish-speaking	0.816	0.311	0.554	31.000
Muslim Teachers	1.048	0.297	0.872	42.000
CNN	0.711	0.133	0.029	206.000

Table A4: Study A (Directs First): Placebo Test II

	Estimate	SE
Nuclear Power	-0.059	0.043
Public Transportation	0.017	0.044
Spanish-speaking	0.031	0.021
Muslim Teachers	-0.009	0.025
CNN	0.075	0.044

Table A5: Study B (Lists First): Three Estimates of Prevalence

	Direct		Standard List		Combined List		% Reduction in Sampling Variance
	$\overline{Y}$	SE	$\hat{\mu}_S$	SE	$\hat{\mu}$	SE	
Nuclear Power	0.591	0.022	0.364	0.087	0.521	0.054	61.047
Public Transportation	0.520	0.022	0.602	0.073	0.672	0.049	54.183
Spanish-speaking	0.099	0.013	0.025	0.079	0.094	0.074	11.822
Muslim Teachers	0.103	0.014	0.195	0.082	0.164	0.074	18.610
CNN	0.457	0.022	0.586	0.107	0.604	0.075	50.158

n = 506 for all estimates

Table A6: Study B (Lists First): Placebo Test I

	$\hat{\beta}$	SE	p-value	n
Nuclear Power	0.739	0.110	0.018	299.000
Public Transportation	0.795	0.097	0.034	263.000
Spanish-speaking	0.301	0.262	0.008	50.000
Muslim Teachers	1.005	0.297	0.987	52.000
CNN	0.759	0.141	0.087	231.000

Table A7: Study B (Lists First): Placebo Test II

	Estimate	SE
Nuclear Power	-0.003	0.044
Public Transportation	0.075	0.045
Spanish-speaking	-0.001	0.027
Muslim Teachers	0.039	0.027
CNN	0.092	0.044

Table A8: Factorial Design: Study A - Study B

	Direct		Conventional List		Combined List	
	Difference	SE	Difference	SE	Difference	SE
Nuclear Power	0.055	0.031	0.210	0.123	0.127	0.073
Public Transportation	0.036	0.031	0.032	0.103	-0.019	0.069
Spanish-speaking	-0.038	0.017	0.093	0.111	0.010	0.104
Muslim Teachers	-0.020	0.018	-0.161	0.113	-0.128	0.104
CNN	-0.049	0.031	-0.330	0.146	-0.306	0.105