Assessing the Validity of Prevalence Estimates in Double List Experiments

Appendix

Gustavo Diaz*

February 13, 2022

A. Experimental design

The main text uses data from a previous published double list experiment (DLE) on support for anti-immigration organizations in California. This was part of a broader study seeking to understand how inattentive respondents behave. See Alvarez et al. (2019b) for details and Alvarez et al. (2019a) for replication materials.

This is a double list experiment with two treatment items, conducted online with a sample of California residents, N = 2725. The study had a total of three attention checks. The DLE appears after the first check. 575 participants failed the first attention check and were dropped before the DLE. That leaves a sample of 2150. Table 1 in the main text reports the distribution of participants across treatment conditions. Among those, only one participant has missing outcomes for one of the DLE questions. This individual is dropped from analyses.

The preamble of the questions is (cf. footnote 10 of main paper):

"Below is a list with the names of different groups and organizations on it. After reading the entire list, we'd like you to tell us how many of these groups and organizations you broadly support, meaning that you generally agree with the principles and goals of the group or organization. Please don't tell us which ones you generally agree with; ONLY TELL US HOW MANY groups or organizations you broadly support. HOW MANY, if any, of these groups and organizations do you broadly support."

*Postdoctoral Fellow. Center for Inter-American Policy and Research. Tulane University. E-mail: gustavodiaz@tulane.edu

Then they observe two baseline lists (cf. Table B7 in appendix):

List A

- Californians for Disability (organization advocating for people with disabilities)
- California National Organization for Women (organization advocating for women's equality and empowerment)
- American Family Association (organization advocating for pro-family values)
- American Red Cross (humanitarian organization)

List B

- American Legion (veterans service organization)
- Equality California (gay and lesbian advocacy organization)
- Tea Party Patriots (conservative group supporting lower taxes and limited government)
- Salvation Army (charitable organization)

List A always appears first. The experiment then includes two sensitive organizations as treatments. The names of the organizations are hidden for ethical reasons.

- Organization X (organization advocating for immigration reduction measures against undocumented immigration)
- Organization Y (citizen border patrol group combating undocumented immigration)

The sensitive items appear randomly in list A or B and are mutually exclusive, so that a respondent that sees X will never see Y. This is why the main text treats them as separate experiments.

I choose this study because using two sensitive items helps illustrate the challenge of creating baseline lists. For organization X, respondents seem to behave as expected, but the pattern of treatment effects shown in Figure 1 of the main text suggests unexpected behaviors for Organization Y.

B. JEPS Reporting Guidelines

This project reanalyzes a previously published experiment. I refer the reader to the original study for details on how the experiment was conducted (Alvarez et al. 2019b, 2019a).

C. APSA's principles and guidance for human subjects research

The original study presented participants with the names of real sensitive organizations, but the published version and replication materials censors them to protect participants and the organizations in question. I never sought to learn nor was made aware of the names of the sensitive organizations. Therefore, any measures taken to protect human subjects in the original study stay the same.

D. Additional results

Application

- Figure D1 shows the mean list experiment outcomes (number of organization that respondent supports) across sensitive items and treatment schedules.
- Tables D1 and D2 compare means across sensitive items and treatment schedules, respectively. Overall, they suggest little evidence against the null hypothesis of equal means, which implies randomization worked as intended.
- Table D3 presents the results of Stephenson's signed rank test for additional subset sizes m. The conclusion does not change.

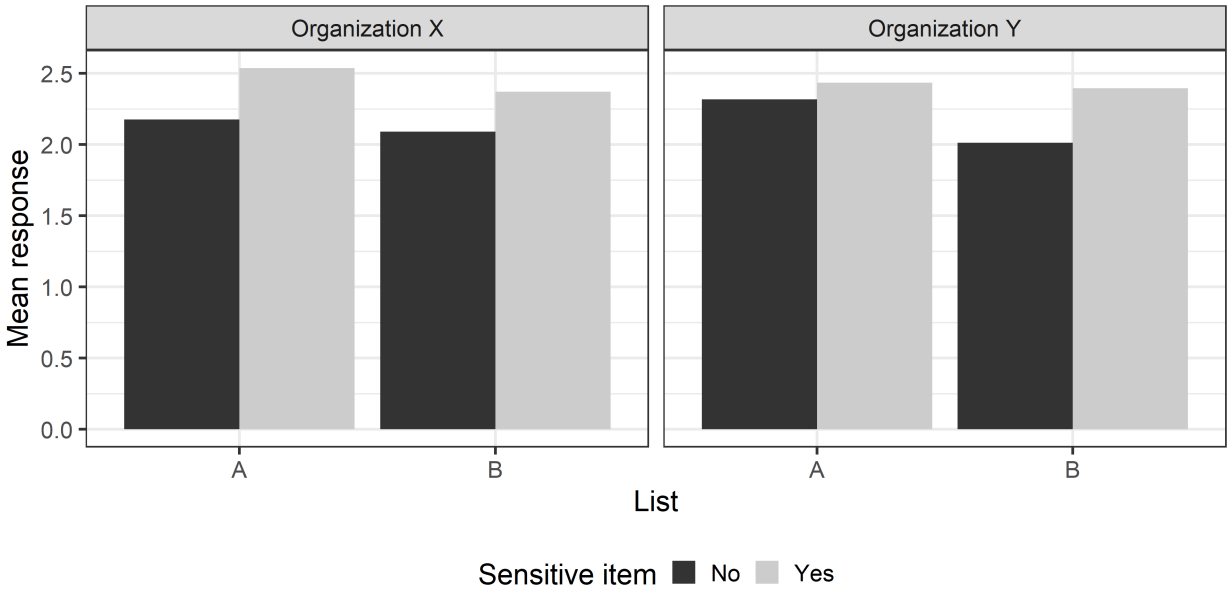

Figure D1: Mean number of organizations respondents support by sensitive item and treatment schedule

Table D1: Comparing means across sensitive items

	Experiment X	Experiment Y	Adj. diff.	Std. diff.	p-value
List B treatment	0.491	0.502	0.011	0.023	0.602
Female	0.533	0.508	-0.025	-0.049	0.255
Age	44.071	43.442	-0.629	-0.038	0.376
No high school	0.028	0.034	0.005	0.030	0.482
High school	0.192	0.203	0.011	0.027	0.527
Some college	0.369	0.381	0.012	0.024	0.576
College	0.299	0.281	-0.018	-0.040	0.352
Post-graduate	0.112	0.102	-0.010	-0.031	0.470
Bay Area	0.169	0.147	-0.022	-0.062	0.155
SoCal (non-LA)	0.308	0.293	-0.016	-0.034	0.432
Los Angeles	0.258	0.282	0.023	0.052	0.227
Central/Southern	0.124	0.138	0.014	0.040	0.353
North/Mountain	0.049	0.061	0.012	0.051	0.241
Central Valley	0.091	0.080	-0.010	-0.037	0.393

Simulation

- Figure D2 shows how increasing the proportion of inflation or deflation induces bias in list experiment estimates. The bias for the list B estimator is more moderate because inflation and deflation move both treatment and control in the same dimension. The bias does not depend on the correlation between lists *rho* by construction since unintended responses happen at random. This does not need to be true in practice, but also not necessary for simulations to be informative.

- Figure D3 shows power for Stephenson's signed rank test for additional subset sizes. Overall, the performance is similar across values of m. This happens because list experiments outcomes have relatively narrow distributions. Still, to avoid cherry-picking, researchers should calibrate and specify a range of m at the pre-analysis stage.

- Figure D4 shows the power of the difference in differences and Stephenson's signed rank ($m = 10$) at fixed deflation rates $\delta = (1, 2, 3)$. The figure shows that increasing deflation rates exacerbate the situation under which the power of the difference in differences decreases with the correlation across lists, whereas the power of signed rank test increases.

Table D2: Comparing means across treatment schedules

	List A	List B	Adj. diff.	Std. diff.	p-value
Experiment Y	0.497	0.509	0.011	0.023	0.602
Female	0.519	0.522	0.003	0.006	0.887
Age	43.280	44.235	0.954	0.058	0.179
No high school	0.028	0.034	0.006	0.035	0.421
High school	0.203	0.192	-0.011	-0.027	0.541
Some college	0.362	0.388	0.027	0.055	0.206
College	0.298	0.281	-0.017	-0.037	0.391
Post-graduate	0.109	0.104	-0.005	-0.017	0.699
Bay Area	0.173	0.143	-0.030	-0.082	0.059
SoCal (non-LA)	0.299	0.302	0.003	0.007	0.879
Los Angeles	0.262	0.278	0.017	0.038	0.385
Central/Southern	0.124	0.137	0.013	0.039	0.374
North/Mountain	0.052	0.058	0.005	0.024	0.583
Central Valley	0.090	0.081	-0.008	-0.030	0.495

Table D3: Stephenson's signed rank test with additional subset sizes for Alvarez et al (2019)

m	Statistic	p-value
Organization X		
2	83.56×10^3	1
5	3.809×10^{12}	1
10	179.2×10^{21}	1
50	143.9×10^{84}	1
Organization Y		
2	35.71×10^3	1
5	3.323×10^{12}	1
10	182.6×10^{21}	1
50	253.4×10^{84}	1

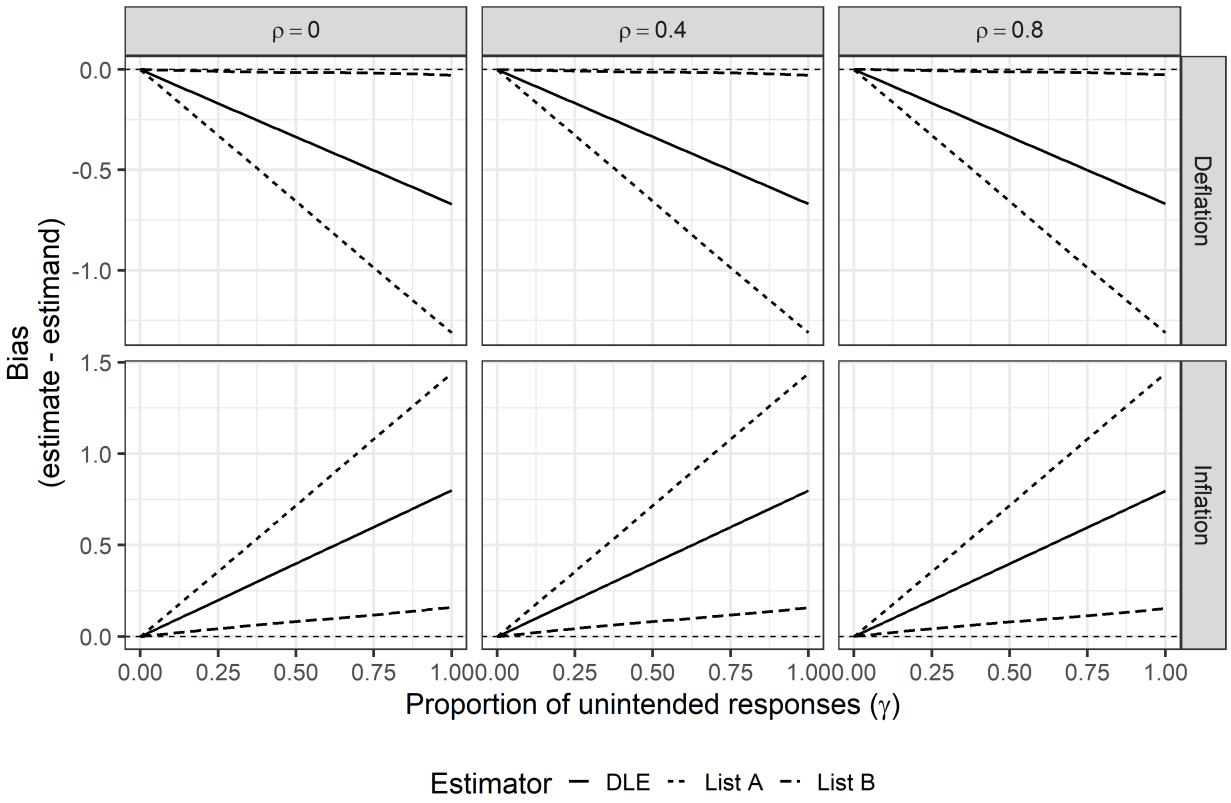

Figure D2: Bias induced by response deflation and inflation across estimators

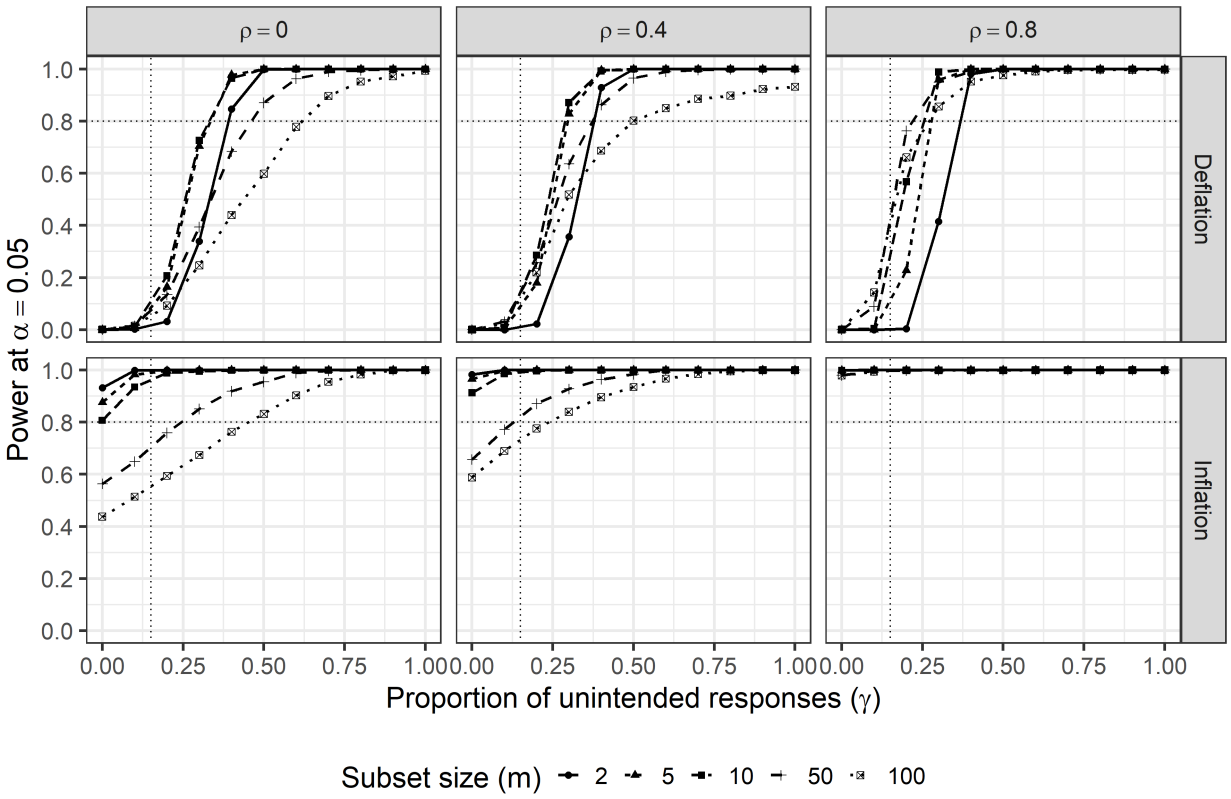

Figure D3: Power of Stephenson's signed rank test under additional subset sizes

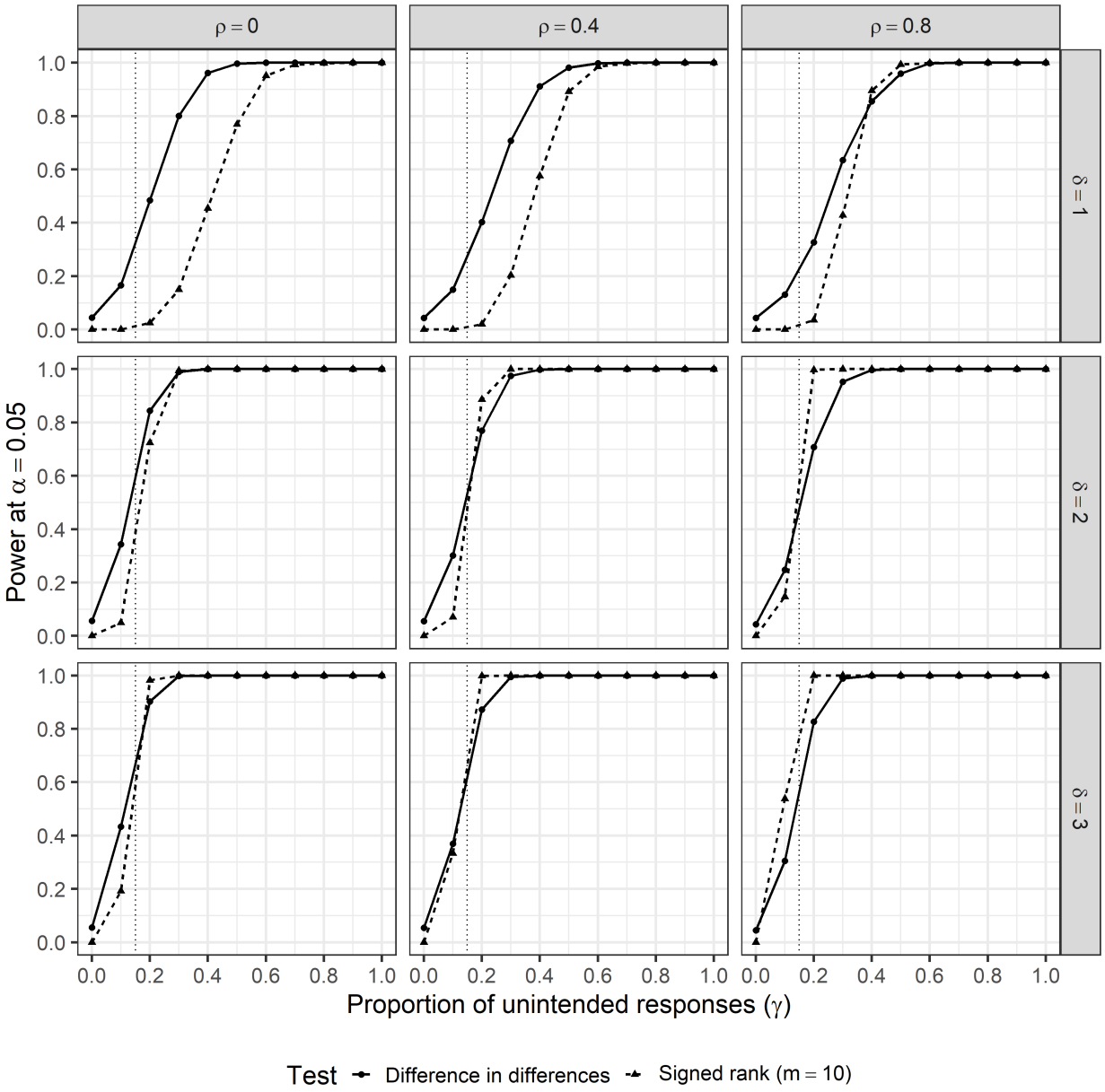

Figure D4: Power at increasing magnitudes of response deflation

E. An illustration of how paired response patterns affect Stephenson's signed rank

Section 3.3 in the main document mentions that only response deflation contributes to a negative Stephenson's signed rank, whereas both inflation and reporting the sensitive item contribute to a positive test statistic.

To further illustrate this point, Table E1 shows a toy example that follows the notation of Section 3. This a single respondent that, as a baseline, would report the same number of items on both questions. Let $z_i = 1$ denote the potential outcomes for each question when the sensitive item appears first and $z_i = 0$ denote potential outcomes when it appears second.

Table E1: An illustration of the behavior of Stephenson's signed rank

Potential outcome	Y_{iA}	Y_{iB}	$(Y_{iA} - Y_{iB})$	$(z_i - (1 - z_i))(Y_{iA} - Y_{iB})$
Baseline	2	2	0	0
Deflation				
$z_i = 1$	1	1	0	0
$z_i = 0$	2	1	1	-1
Inflation				
$z_i = 1$	3	3	0	0
$z_i = 0$	2	3	-1	1
Sensitive item				
$z_i = 1$	3	2	1	1
$z_i = 0$	2	3	-1	1

The table considers three scenarios:

- Response deflation of one unit with a zero prevalence sensitive item (to facilitate interpretation).
- Response inflation of one unit.
- Reporting the sensitive item without response deflation nor inflation under no liars and no design effects.

For each scenario, one can calculate $(Y_{iA} - Y_{iB})$ and $(z_i - (1 - z_i))(Y_{iA} - Y_{iB})$. This last quantity indicates how the observed pattern of responses contributes to the test statistic.

The table highlights three results:

- Deflation contributes negatively to the test statistic only when the sensitive item appears second.
- Inflation contributes positively only when the sensitive item appears second.
- Reporting the sensitive item contributes positively regardless of the placement of the sensitive item.

In practice, the contribution varies with the values of Y_{iA} and Y_{iB}, but this reflects the intuition of why the test performs well at detecting deflation but struggles with inflation.

References

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019a. "Replication Data for: Paying Attention to Inattentive Survey Respondents." Harvard Dataverse. https://doi.org/10.7910/DVN/TUUYLQ.

———. 2019b. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27 (2): 145–62.