# Assessing the Validity of Prevalence Estimates in Double List Experiments

# Appendix

Gustavo Diaz*

## Contents

**A. Experimental design** **2**

**B. JEPS Reporting Guidelines** **3**

**C. APSA's principles and guidance for human subjects research** **3**

**D. Additional results** **3**

Application . . . . . 3

Simulation . . . . . 4

**E. The cost of implementing a double list experiment** **10**

**References** **11**

*Postdoctoral Fellow. Department of Political Science. McMaster University. E-mail: diazg2@mcmaster.ca

# 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 in this paper.

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."

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 to 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.
- Figure D2 shows the distribution of responses in the control condition (no sensitive item) across organizations and baseline lists. The similar distributions across organizations suggest that ceiling or floor effects are an unlikely explanation for the pattern of estimates observed in Figure 1 in the main text.

- 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.

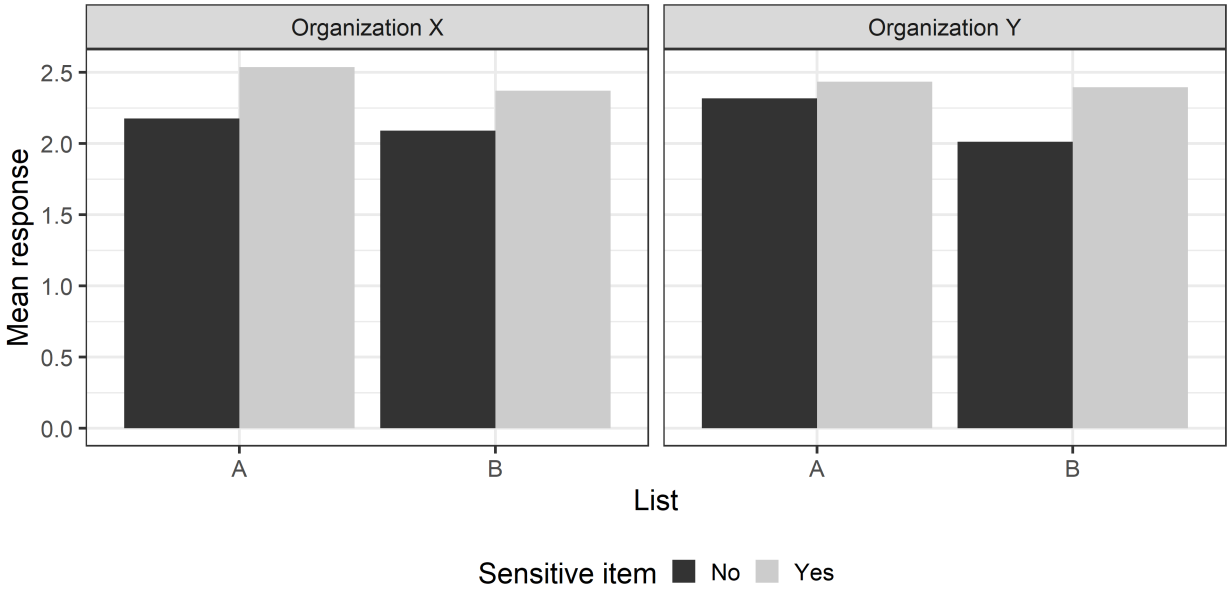


**Figure D1: Mean number of organizations respondents support by sensitive item and treatment schedule**

## Simulation

- Figure D3 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 D4 shows power simulations 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. The figure is a useful example of how to do so.
- Figure D5 shows the power of the difference in differences and Stephenson's signed rank ($m = 10$)

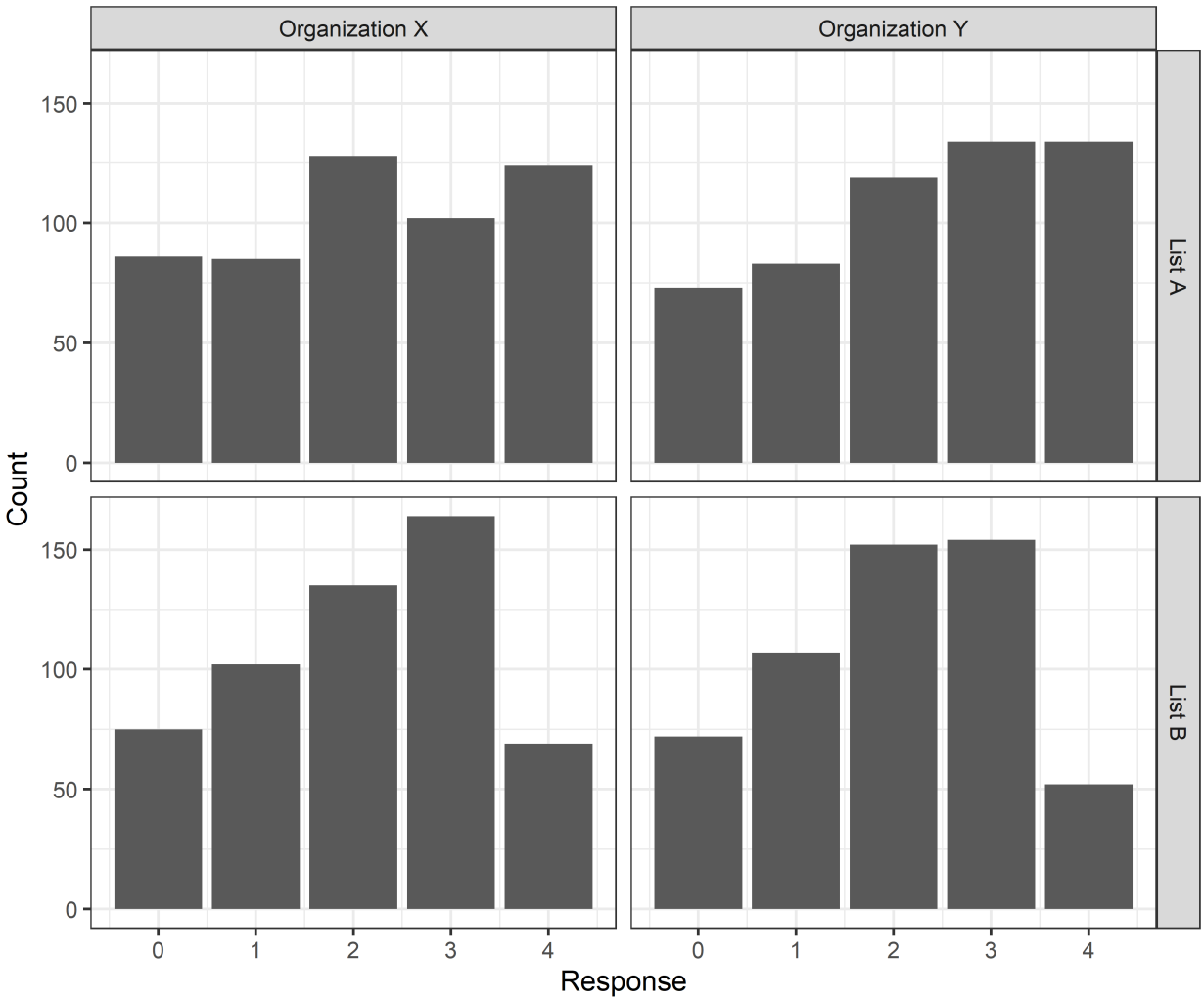


**Figure D2: Distribution to responses in the control condition across organizations and baseline lists**

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 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 |

**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 | $83.56 \times 10^3$ | 1 |
| 5 | $3.809 \times 10^{12}$ | 1 |
| 10 | $179.2 \times 10^{21}$ | 1 |
| 50 | $143.9 \times 10^{84}$ | 1 |
| **Organization Y** | | |
| 2 | $35.71 \times 10^3$ | 1 |
| 5 | $3.323 \times 10^{12}$ | 1 |
| 10 | $182.6 \times 10^{21}$ | 1 |
| 50 | $253.4 \times 10^{84}$ | 1 |

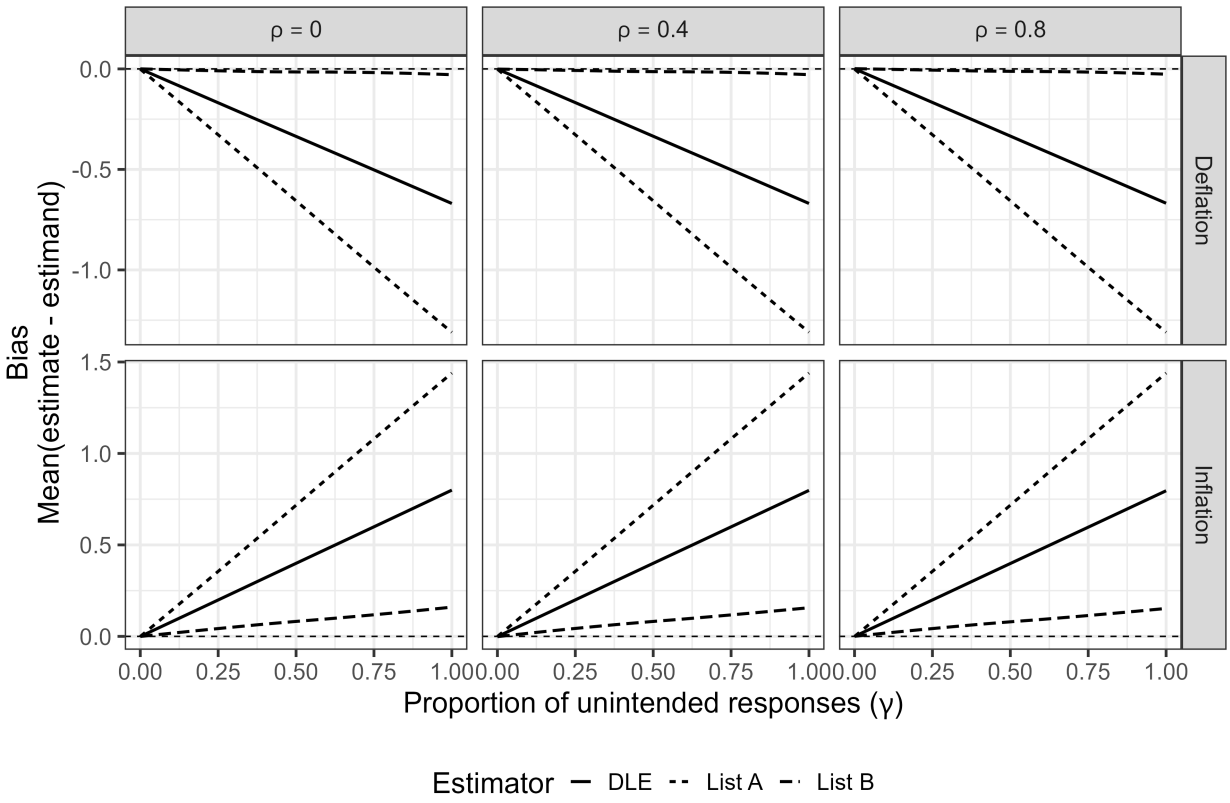


**Figure D3: Bias induced by response deflation and inflation across estimators**

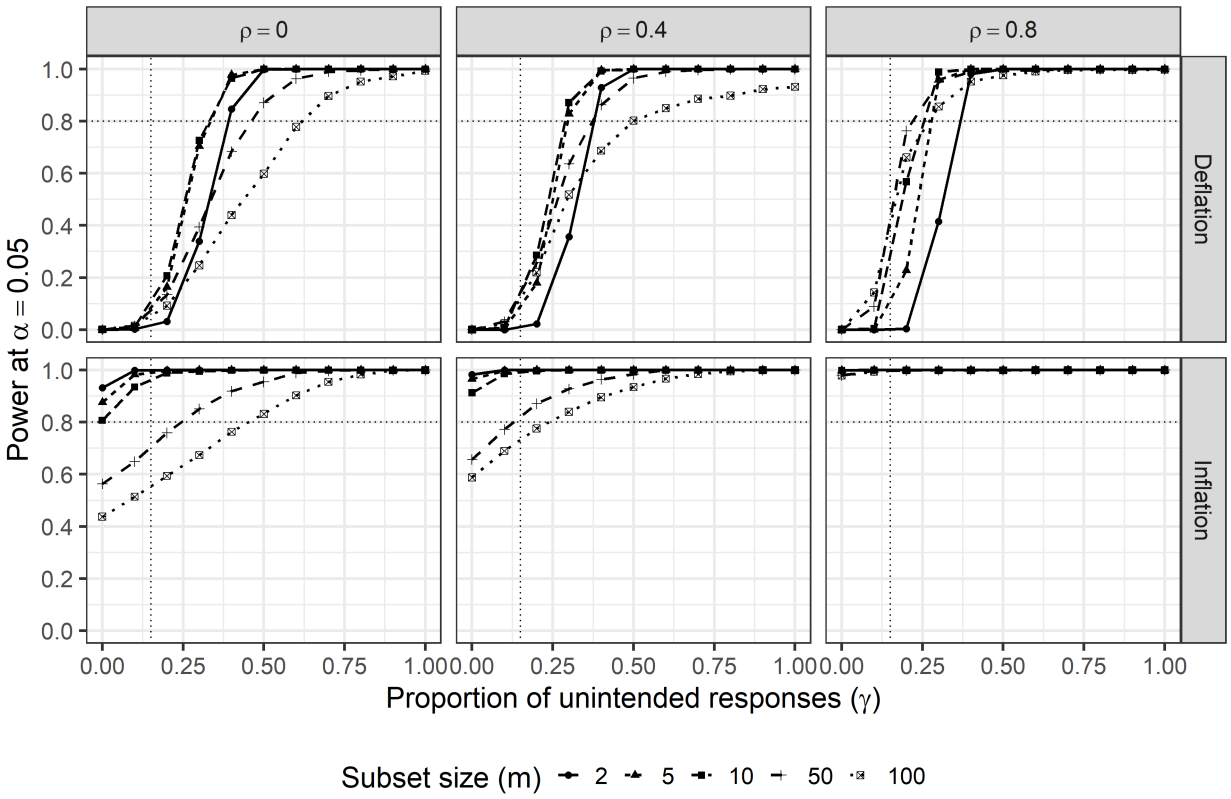


**Figure D4: Power of Stephenson's signed rank test under additional subset sizes**

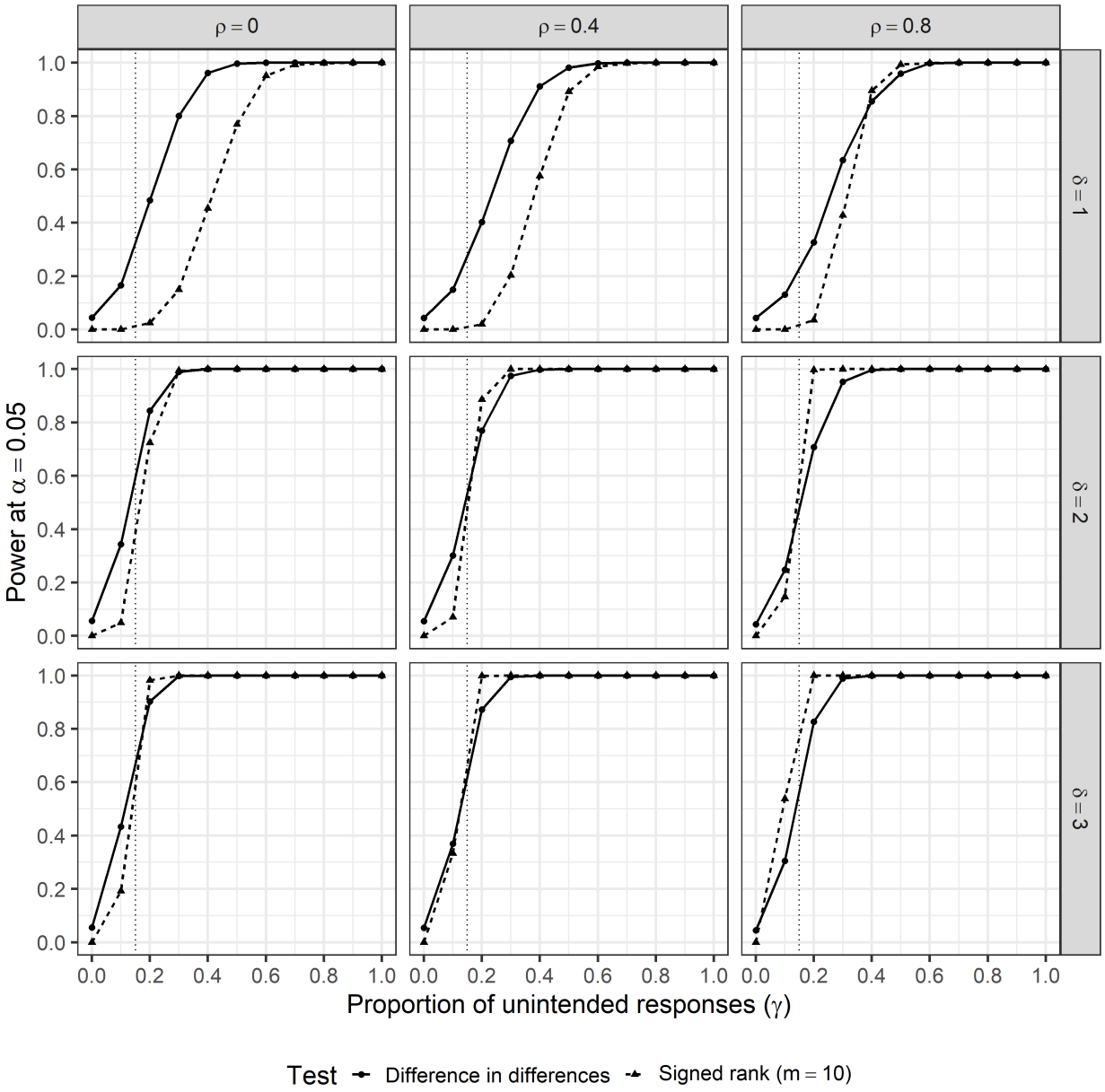


**Figure D5: Power at increasing magnitudes of response deflation**

## E. The cost of implementing a double list experiment

The main text mentions that in some cases the additional cost of implementing a DLE may offset the improvement in precision. This is because the extensive piloting required to validate two baseline lists may limit the effective budget to conduct confirmatory analysis.

This section illustrates how to navigate this cost-benefit calculation with simulations. I simulate single and double list experiments following the same parameters than Section 4 in the main text, except that respondents do not engage in any strategic misreporting.

I consider the rank correlation between lists $\rho = \{0, 0.4, 0.8\}$. I represent the cost of implementing a DLE as a function of the proportion of the potential sample that would be lost compared to a single list experiment. A researcher would have a smaller sample if, for example, a DLE requires additional rounds of pre-testing. For simplicity, I assume single list experiments do not incur sample loss.

Figure E1 shows the power of single and double list experiment estimators over 1,000 simulations at each parameter combination. The figure shows that single list experiments have about 60% power to detect a non-zero prevalence effect. This is constant with increasing sample loss by design. Double list experiments start with power over 80% and decrease with increasing sample loss. The decrease is less pronounced when the correlation between the two baseline lists increases because the estimator is more precise (Glynn 2013).

Overall, the figure suggests that a researcher should be indifferent between well designed single and double list experiments if they anticipate no correlation between baseline lists and losing about half of the potential sample by implementing a DLE. While this suggests DLEs are generally advisable, the final decision should be made based on simulations that account for researchers' knowledge of the context under study.

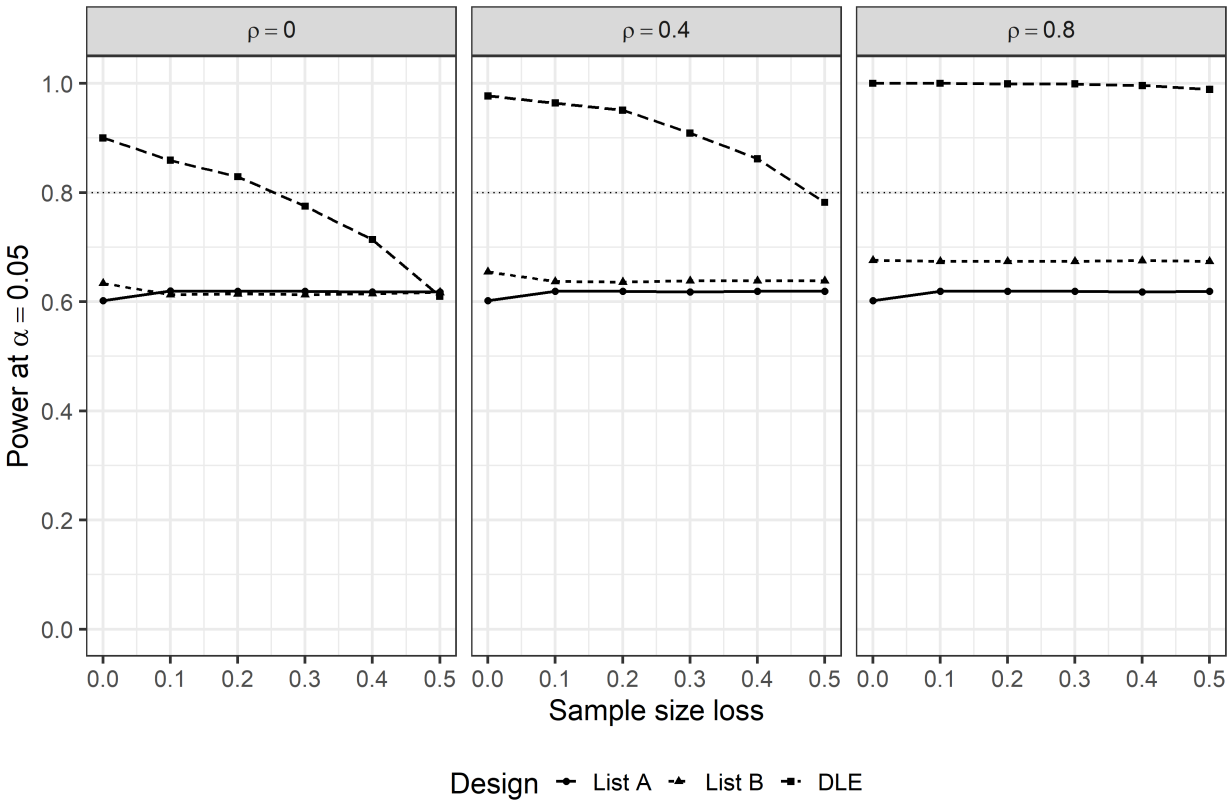


**Figure E1: Cost of implementing a DLE as a function of sample loss**

# 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.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum?" *Public Opinion Quarterly* 77 (S1): 159–72.