Recovering Informative Estimates in Failed Placebo-Controlled List Experiments

Supplementary Materials

Gustavo Díaz* Ines Fynn† Verónica Pérez Bentancur‡

Lucía Tiscornia§

June 10, 2026

Table of contents

*Assistant Professor of Instruction. Northwestern University. Email: gustavo.diaz@northwestern.edu

†Assistant Professor, Department of Social Science, Universidad Católica del Uruguay, ines.fynn@ucu.edu.uy

‡Assistant Professor, Department of Political Science, Universidad de la República, veronica.perez@cienciassociales.edu.uy

§Associate Professor, School of Politics and International Relations, University College Dublin, lucia.tiscornia@ucd.ie

A. Frequency of placebo controlled list experiments

The main text mentions how placebo-controlled list experiments are used infrequently despite their promise to reduce bias from non-strategic errors in a cost-efficient manner. Here we present some data to support that claim based on an AI-assisted snowball search using Claude Sonnet 4.6. All AI-generated output was manually checked by the researchers and we take full responsibility for any mistakes.

We started with two seed lists of previously conducted list experiments (Imai 2011; Blair et al. 2020). The list contains both published and unpublished work using list experiments in political science and adjacent journals. We updated the entries for manuscripts that were coded as unpublished but became published after the original lists were created. We then conducted forward searches within the same 48 journals already included in the list for new entries that include list experiments, focusing on the 2020-2025 range since the entries in the original list go up to 2019.

At this point, we noticed that the list omitted Riambau and Ostwald (2020) and Agerberg and Tannenberg (2021), two recent papers that make an explicit case for placebo-control design variants. We included these in the list and augmented it with the list of published articles publicly available unpublished manuscripts that directly cited them up to the year 2026.

After filtering out false positives (e.g. studies that mentions list experiments but do no report the results of one), we ended up with a list of 120 unique manuscripts. We then coded whether each entry uses a placebo item or a placebo control design. The distinction is important because some studies mention using a placebo item for validation but do not resort to the specific placebo-controlled design discussed in the main text (e.g. Ahlquist et al. 2014).

The first mention of placebo items dates back to Ahlquist et al. (2014), whereas Riambau and Ostwald (2020) is the first published entry that explicitly suggests a placebo-controlled design

(although their application does not use such a design). Out of the 81 entries since 2014, ten use a placebo item of some sort ($\approx 12.4\%$), while only five of them use a placebo-controlled design ($\approx 6.17\%$), with the oldest entries starting in the year 2021.

This is far from being a comprehensive or representative list, but it is sufficient to sustain the claim that this research design innovation has not become common practice in political science and adjacent fields. We do not have enough information to determine the reason for the infrequent use, but this aligns with previous work showing that empirical work using list experiments rarely deviates from the conventional design (Blair et al. 2020).

B. Descriptive statistics

- Table B1 shows descriptive statistics on available demographic variables.
- Figure B1 shows the distribution of responses to both list A and B, divided by treatment (includes sensitive item) and control group (placebo item). Overall, the distribution of responses and associated χ^2 tests do not show evidence of treatment responses deflating in large magnitude. Across panels, we do not enough have evidence against the null hypothesis of treatment and control responses being drawn from different distributions. The exception is the third panel on the right ($p = 0.039$). However, the Bonferroni-corrected threshold for statistical significance at conventional levels is $0.05/8 \approx 0.006$ since we are performing multiple comparisons.

Table B1: Sample characteristics

Variable	N	Mean	SD	Min	Max
Age (years)	2668	41.372	15.192	18	100
Sex					
Female	2668	0.642	0.479	0	1
Male	2668	0.353	0.478	0	1
Other sex	2668	0.005	0.070	0	1
Education					
Primary incomplete	2555	0.023	0.149	0	1
Primary complete	2555	0.069	0.254	0	1
Secondary incomplete	2555	0.291	0.454	0	1
Secondary complete	2555	0.190	0.392	0	1
Tertiary incomplete	2555	0.264	0.441	0	1
Tertiary complete	2555	0.163	0.370	0	1
Race					
White	2525	0.741	0.438	0	1
Afro-descendant	2525	0.104	0.306	0	1
Indigenous	2525	0.069	0.253	0	1
Other race	2525	0.086	0.280	0	1
Employment					
Worked last week	2545	0.560	0.496	0	1
Income					
Not enough, faces great difficulties	2519	0.112	0.316	0	1
Not enough, faces some difficulty	2519	0.353	0.478	0	1
Enough to make ends meet	2519	0.424	0.494	0	1
Enough to save	2519	0.110	0.313	0	1

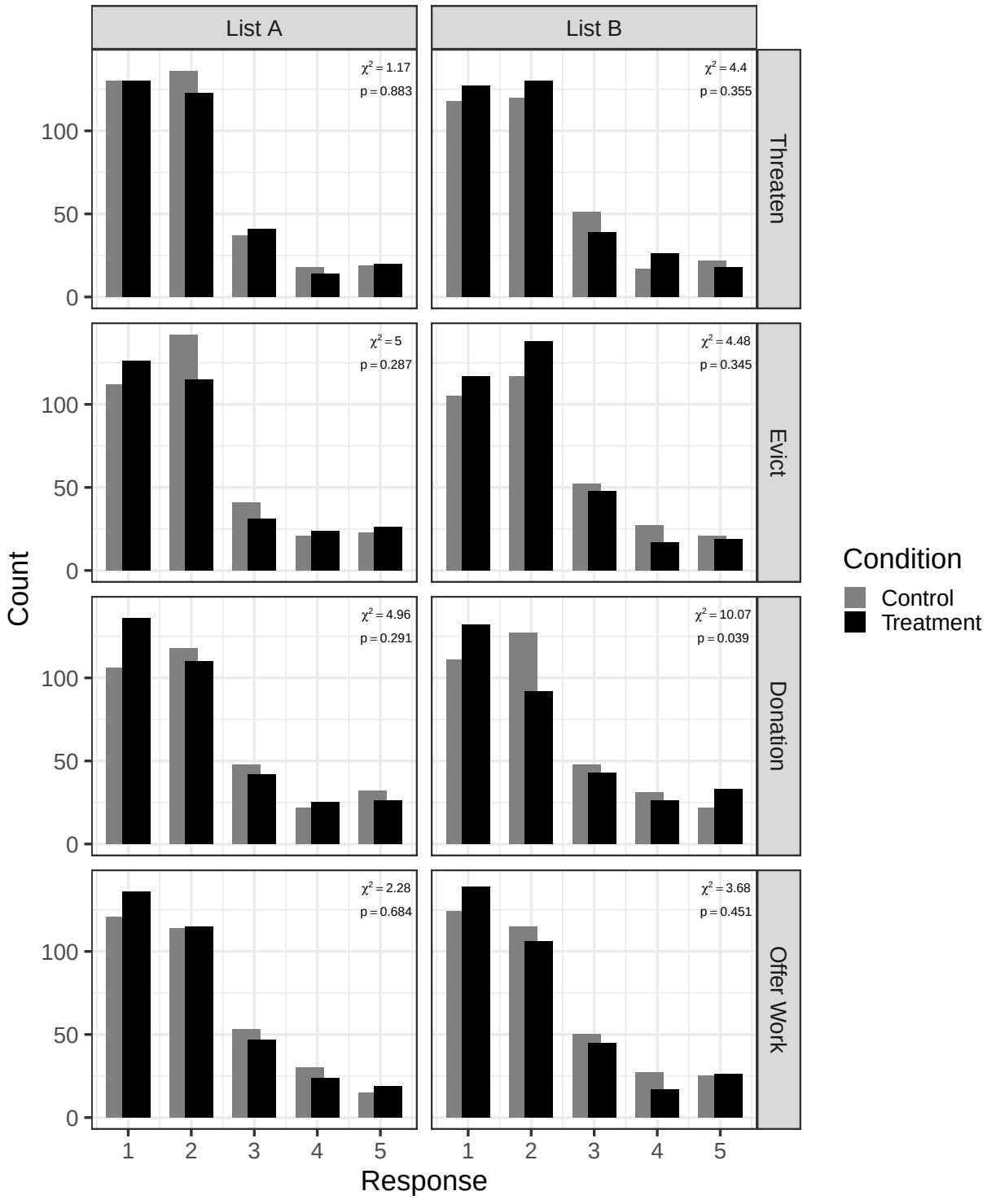

Figure B1: Distribution on responses to list experiment questions with χ^2 tests

C. Additional results

- Table C1 shows the interaction coefficient between failing the mock vignette attention check and the sensitive item placement indicator. This can be interpreted as an estimate on whether failing the attention check alters treatment effect responses.
- Table C2 shows the interaction coefficient between the raw mock vignette responses and the sensitive item placement indicator.
- Table C3 shows the interaction coefficient between the numerical distance from the correct answer in the mock vignette recall question and the sensitive item placement indicator
- Table C4 shows OLS regression results for the relationship between the mock vignette responses and total time spent on list experiment responses

Table C1: Effects do not vary by failed mock vignette attention check

Item	Estimate	SE	p-value
List A			
Threatening neighbors	0.015	0.346	0.96495
Evicting neighbors	0.132	0.268	0.6239
Making donations	-0.253	0.233	0.28112
Offering work	0.335	0.196	0.091
List B			
Threatening neighbors	-0.015	0.475	0.97478
Evicting neighbors	0.402	0.411	0.32851
Making donations	-0.615	0.402	0.12648
Offering work	-0.244	0.325	0.45411
Double			
Threatening neighbors	0.128	0.456	0.77951
Evicting neighbors	-0.192	0.434	0.65909
Making donations	0.140	0.454	0.75733
Offering work	0.935	0.335	0.00539

Note: Estimates represent the regression coefficient for the interaction between a binary indicator denoting whether the respondent failed the attention check and the sensitive item placement indicator. Drawn from OLS regression models with block fixed effects and HC2 robust standard errors (for List A and) and clustered standard errors by respondent (for the Double list experiment).

Table C2: Effects do not vary by raw mock vignette response

Item	Estimate	SE	p-value
List A			
Threatening neighbors	-0.005	0.046	0.9137
Evicting neighbors	-0.002	0.039	0.9515
Making donations	0.005	0.037	0.8991
Offering work	-0.033	0.034	0.3424
List B			
Threatening neighbors	-0.043	0.067	0.517
Evicting neighbors	0.069	0.065	0.2858
Making donations	-0.108	0.068	0.1162
Offering work	-0.097	0.057	0.0879
Double			
Threatening neighbors	0.036	0.066	0.5869
Evicting neighbors	-0.073	0.064	0.2507
Making donations	0.122	0.071	0.0844
Offering work	0.035	0.059	0.558

Note: Estimates represent the regression coefficient for the interaction between the raw mock vignette response and the sensitive item placement indicator. Drawn from OLS regression models with block fixed effects and HC2 robust standard errors (for List A and List B) and clustered standard errors by respondent (for the Double list experiment).

Table C3: Effects do not vary by distance from correct answer on recalling the number of items on mock vignette

Item	Estimate	SE	p-value
List A			
Threatening neighbors	0.014	0.044	0.746
Evicting neighbors	0.004	0.039	0.924
Making donations	-0.020	0.037	0.595
Offering work	-0.030	0.041	0.465
List B			
Threatening neighbors	0.096	0.075	0.199
Evicting neighbors	-0.058	0.076	0.445
Making donations	0.049	0.079	0.533
Offering work	-0.059	0.073	0.424
Double			
Threatening neighbors	-0.068	0.075	0.366
Evicting neighbors	0.056	0.068	0.414
Making donations	-0.087	0.080	0.278
Offering work	-0.007	0.076	0.932

Note: Estimates represent the regression coefficient for the interaction between the distance from the correct answer on recalling the number of items on the mock vignette and the sensitive item placement indicator. Drawn from OLS regression models with block fixed effects and HC2 robust standard errors (for List A and List B) and clustered standard errors by respondent (for the Double list experiment).

Table C4: Mock vignette responses and time spent on list experiment questions

	Attention check	Raw response
Intercept	72.096*	71.169*
	(0.787)	(1.446)
Failed attention check	0.571	
	(2.315)	
Raw response		0.369
		(0.418)
Num.Obs.	2668	2668
R2	0.000	0.000

* p <0.05

Note: OLS regression models with HC2 robust standard errors. Outcome is total time spent on both list experiment questions (in seconds).

References

Agerberg, Mattias, and Marcus Tannenberg. 2021. "Dealing with Measurement Error in List Experiments: Choosing the Right Control List Design." *Research & Politics* 8 (2): 205316802110131. https://doi.org/10.1177/20531680211013154.

Ahlquist, John S., Kenneth R. Mayer, and Simon Jackman. 2014. "Alien Abduction and Voter Impersonation in the 2012 u.s. General Election: Evidence from a Survey List Experiment." *Election Law Journal: Rules, Politics, and Policy* 13 (4): 460–75. https://doi.org/10.1089/elj.2013.0231.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114 (4): 1297–315.

Imai, Kosuke. 2011. *Examples of List Experiments*. https://imai.fas.harvard.edu/research/files/listExamples.pdf.

Riambau, Guillem, and Kai Ostwald. 2020. "Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore." *Political Science Research and Methods* 9 (1): 172–79.