Recovering Informative Estimates in Failed Placebo-Controlled List Experiments*

Early draft – Ask before citing

Gustavo Díaz† Ines Fynn‡ Verónica Pérez Bentancur§

Lucía Tiscornia¶

March 2, 2026

Abstract

List experiments are a popular technique to reduce sensitivity bias in surveys. Their widespread use has spurred considerable research design innovation. One suggestion to minimize bias from respondent non-strategic errors is to include a placebo statement in the control list. However, a poorly chosen placebo may introduce further bias, to a point in which estimates are no longer valid. This paper shows how to recover estimates in failed placebo-controlled list experiments. We present an example of a (double) list experiment on exposure to criminal governance tools in Uruguay in which the inclusion of a placebo item leads to invalid prevalence estimates. We then show how to leverage experimental assumptions and auxiliary information to recover informative estimates via partial identification bounds.

Word count: 3803 (including references).

*We thank participants at MAPOR 2025 for feedback.

†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

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

1 Introduction

List experiments are a popular technique to reduce sensitivity bias in surveys (Droitcour et al. 1991). Their widespread use in the discipline has spurred innovations aimed at establishing best research design practices (See Blair, Coppock, and Moor 2020 for a review).[1]

Some of these innovations seek to minimize *non-strategic* response errors. For example, inattentive respondents choosing alternatives at random or straight-lining may create an artificial correlation between treatment assignment and observed responses just because treatment group lists are longer, which inflates prevalence rate estimates. In this case, the solution would be to include a zero-prevalence placebo statement in (at least some of) the control group so that list length is constant across experimental conditions (Riambau and Ostwald 2020; Agerberg and Tannenberg 2021).[2]

In this paper, we focus on the problem that emerges when the inclusion of a placebo statement backfires. In theory, one should include a placebo item in the control group that has (near) zero prevalence so that it does not induce further bias in the estimation of sensitive item prevalence rates. In practice, researchers may inadvertently choose a placebo item to which a non-trivial number of participants respond positively, either because the researcher underestimates its prevalence or because the placebo item stands out in a way that induces unintended response patterns. In this case, a failed placebo-controlled list experiment attenuates prevalence rate estimates.

This paper shows how to recover informative estimates in failed placebo-controlled list experiments using a partial identification approach. We first present an example of a list experiment on the prevalence of exposure to criminal governance tools in Uruguay in which the inclusion of a placebo induces prevalence rate estimates that are essentially zero. The

[1] Another strand of the literature seeks to improve list experiments by developing new tools for estimation (e.g. Blair and Imai 2012; Blair, Imai, and Lyall 2014), which is beyond the scope of this paper.

[2] Previous work also addresses *strategic* response errors. See Ahlquist (2017) and Blair, Chou, and Imai (2019) for a discussion.

nature of our research design, plus additional testing, suggests this attenuation bias emerges from the inclusion of a failed placebo item. We then show how to leverage list experimental assumptions and auxiliary information to produce informative estimate bounds. We conclude with recommendations to safeguard against implementation issues in placebo-controlled list experiments.

2 A (Double) List Experiment with a Placebo Statement

2.1 Research Design

We conducted an online survey on a sample of 2,688 residents in the city of Montevideo, Uruguay in 2022.[3] The main substantive goal of this survey was to document the prevalence of criminal governance actions in a context of high state capacity. Current research focuses mostly on cases characterized by more complex and violent versions of criminal governance, such as Mexico, Brazil, Colombia, or El Salvador (Magaloni, Franco-Vivanco, and Melo 2020; Arias 2017; Barnes 2021). Although at a smaller scale, criminal governance is still a persistent problems in other areas of the region, such as Argentina, Chile, and Uruguay (Auyero and Sobering 2019; Flom 2022; Fynn, Pérez Bentancur, and Tiscornia 2024). Understanding the prevalence of criminal governance in these less studied cases becomes crucial to determine whether mainstream theories apply beyond their origin.

We included list experiment questions in the survey. Participants were presented with two lists of four things people may have experienced in the last six months, and were instructed to indicate how many apply to them. Table 1 shows the two baseline lists, which were presented in random order.

[3]Subjects were recruited via Facebook ads to balance cost with the goal of accessing respondents in areas where criminal governance is prevalent, wherein polling firms may be wary to field in-person surveys.

Table 1: Double list experiment baseline lists

List A	List B
Saw people doing sports	Saw people playing soccer
Visited friends	Chatted with friends
Attended activities by feminist groups	Attended activities by LGBTQ groups
Went to church	Went to charity events

Respondents were asked to indicate how many things they have experience in the last six months for each list. Baseline lists were presented in random order, items within lists were randomly shuffled.

We crafted the baseline lists following standard advice in the literature to avoid ceiling and floor effects, which entails inducing negative correlation between at least two items within each list and positive correlation across lists (Glynn 2013). For example, we expected people who attended activities by feminist groups to be unlikely to also go to church. Following the same line of advice, baseline lists are also constructed so that each item in list A has a similar item in list B, which leads us to expect respondents to give similar answers to both lists of baseline items.

Respondents were randomly assigned to see one of four sensitive items representing both negative and positive criminal governance actions:

1. Saw criminal groups threatening neighbors
2. Saw criminal groups evicting neighbors from their homes
3. Saw criminal groups making donations to neighbors
4. Saw criminal groups offering work to neighbors

In turn, the selected sensitive item would be randomly assigned to appear on List A or List B. Whichever list did not display the sensitive item would show the following placebo item:

- I did not drink *mate*

Mate is a popular herbal infusion in many South American countries. We anticipated this would be a good placebo item as it is tantamount to asking in other contexts whether

Table 2: Sample sizes by experimental condition

Item location		Sensitive item			
Sensitive	Placebo	Threaten	Evict	Donation	Work
List A	List B	328	322	339	341
List B	List A	340	339	326	333

someone has not drank coffee or tea in the last six months. Table 2 shows a breakdown of sample sizes by experimental condition.[4]

We randomized which sensitive item respondents see, the list in which it appears, and the order in which lists are presented based on self-reported neighborhood of residence within Montevideo.[5] We formed two groups based on Montevideo's police classification of neighborhoods as peripheral and non-peripheral zones, and then conducted all randomization procedures independently within each block.

Later in the survey, we asked about each sensitive item directly as a yes/no question as part of a longer battery. This lets us can compare list experiment prevalence estimates with proportions estimated by direct questions. Crucially, we did not ask about the placebo item directly.

2.2 Results

As Table 2 suggests, we have what is equivalent to four independent double list experiments, or eight independent conventional "single" list experiments. We also can also calculate the proportion of respondents who admitted to each sensitive item in a direct question.

Figure 1 shows our estimates for all these quantities of interest. Each panel displays estimates for each sensitive item of interest, while each row indicates the estimator that corresponds to how the question was asked. Across items, the direct question suggest a positive, non-zero

[4]The placement of sensitive estimates was block-randomized based on two neighborhood groups identified by Montevideo's police as low and high crime risk. All results presented include block fixed-effects.

[5]Respondents who did not provide a valid response to this question were dropped from the survey before random assignment to conditions.

prevalence estimate that hovers around 10-20%. As usual, these estimates are much narrower confidence intervals relative to the list experiment questions, but if these are indeed sensitive questions, one would expect point estimates to under-report the true prevalence rate.

Assuming the list experiment work as intended, one would expect it to yield prevalence estimates equal or higher than those drawn from a direct question. In this case, point estimates are consistently around or below zero, in some cases with effect sizes implying -10% prevalence rates. In the context of a double list experiment, differences between single list estimates would hint at potential questionnaire design issues (Diaz 2023), but in this case point estimates in both lists exhibit the same pattern. Moreover, that the pattern is consistent across both positive and negative forms of criminal governance suggests that the problem goes beyond a violation of standard list experimental assumptions.

We argue that the source of this unexpected pattern in estimates is the inclusion of a placebo sensitive item in the control lists. In the following sub-section, we present several statistical tests to rule out alternative explanations.

2.3 Ruling out alternative explanations

Beyond randomization and SUTVA, list experiments require additional assumptions to yield valid estimates of the prevalence of the sensitive item of interest (Blair and Imai 2012; Aronow et al. 2015). The "no liars" assumption requires respondents to not report holding the sensitive item when they do not. This is referred to as "one-sided lying" in the broader indirect questioning literature (Gingerich et al. 2016). This assumption can be violated inadvertently if the putative sensitive item is perceived as socially *desirable* by some subgroups in the study sample [self cite forthcoming].

A second assumption is "no design effects." It requires that responses to baseline lists are not affected by the inclusion of the sensitive item. In the standard list experiment design, this may happen when the inclusion of a sensitive item changes the interpretation of the list

as a whole. In double list experiments, the inclusion of the sensitive item in the first list may inadvertently influence responses in the second list, producing *carryover* design effects (Diaz 2023).

One alternative interpretation for the pattern of prevalence estimates observed in Figure 1 is that the design of the list experiment questions induced response patterns that violate these assumptions. For example, if the sensitive items stand out relative to the baseline lists and they are recognized as socially undesirable, respondents may have deflated their response when a sensitive item is present in an attempt to reduce the probability of being associated with criminal governance activities. This would induce attenuation bias in prevalence estimates, which in an extreme case could also produce Figure 1.

While no liars and no design effects remain assumptions, the literature has developed several tests to detect response patterns that would suggest these assumptions are not justifiable. Table 3 reports the results of three different statistical tests, grouped by sensitive item and baseline list where appropriate:

1. The prevalence estimate among those who admit to the sensitive item in the direct question (Placebo I in Aronow et al. 2015). The original null hypothesis in this test is that prevalence equals one, but because our placebo item appears to offset prevalence estimates, we instead test the null hypothesis of prevalence being equal to zero among in this group of respondents.

2. The effect of the placement of the sensitive item on responses to the direct question (Placebo II in Aronow et al. 2015). The original tests evaluates whether the *presence* of the sensitive item shapes responses to the direct question. In our case, because every respondent sees the sensitive item in one of the lists, we are instead testing whether a particular combination of items affects how respondents engage with the direct question.

3. The effect of placing the sensitive item on the first list on responses to the second

list (Diaz 2023). Because the order of lists A and B are shuffled, this a direct test of carryover design effects.

We opt for these tests as they have straightforward regression analogues that can account for the block-randomized structure of our list experiment and the use of a placebo statement. Other alternatives, including those presented in Blair and Imai (2012) and Glynn (2013), are not directly transportable under deviations in the canonical design.

The tests in Table 3 show little evidence against the the corresponding null hypotheses. Regarding Placebo I, we only find considerable evidence against non-zero prevalence among those who admitting to the sensitive item for seeing criminal groups making donations to neighbors. In this case, prevalence is around 0.46 in list A, but not in list B or their combination. Moreover, for every sensitive item except seeing criminal groups threatening neighbors, the observed prevalence among those admitting to the sensitive item are signed-flipped between list A and B, resulting in a double list prevalence somewhat close to zero. This may imply a form of strategic error, but the small sample size limits our ability to draw inferences.

Moving to Placebo I and the test for carryover design effects, we find little evidence against the corresponding null hypotheses, and the sample sizes suggest that this is not due to low statistical power.

Overall, we find little evidence against strategic response errors. This does not imply that strategic errors do not exist at all, just that it is unlikely that the pattern observed in Figure 1 responds primarily to strategic violations of the standard list experiment assumptions.

Another alternative source for the unusual pattern of prevalence estimates may be response inattention. List experiment questions are more involved tasks than a typical survey item, so respondents are more likely to speed through them or make mistakes. In either case, this can result in non-zero estimates for zero prevalence items. For example, Ahlquist, Mayer, and Jackman (2014) document small but positive prevalence of alien abductions in a list

experiment conducted on a representative sample in the United States.

We used several questions and checks to detect low quality responses due to inattention. Appendix XX presents the results of these checks and shows how the unusual pattern of prevalence estimates exists even after accounting for inattention.[6]

[6]The short version is that we have a very easy screener and a very hard one, the pattern is still there whichever filter we use.

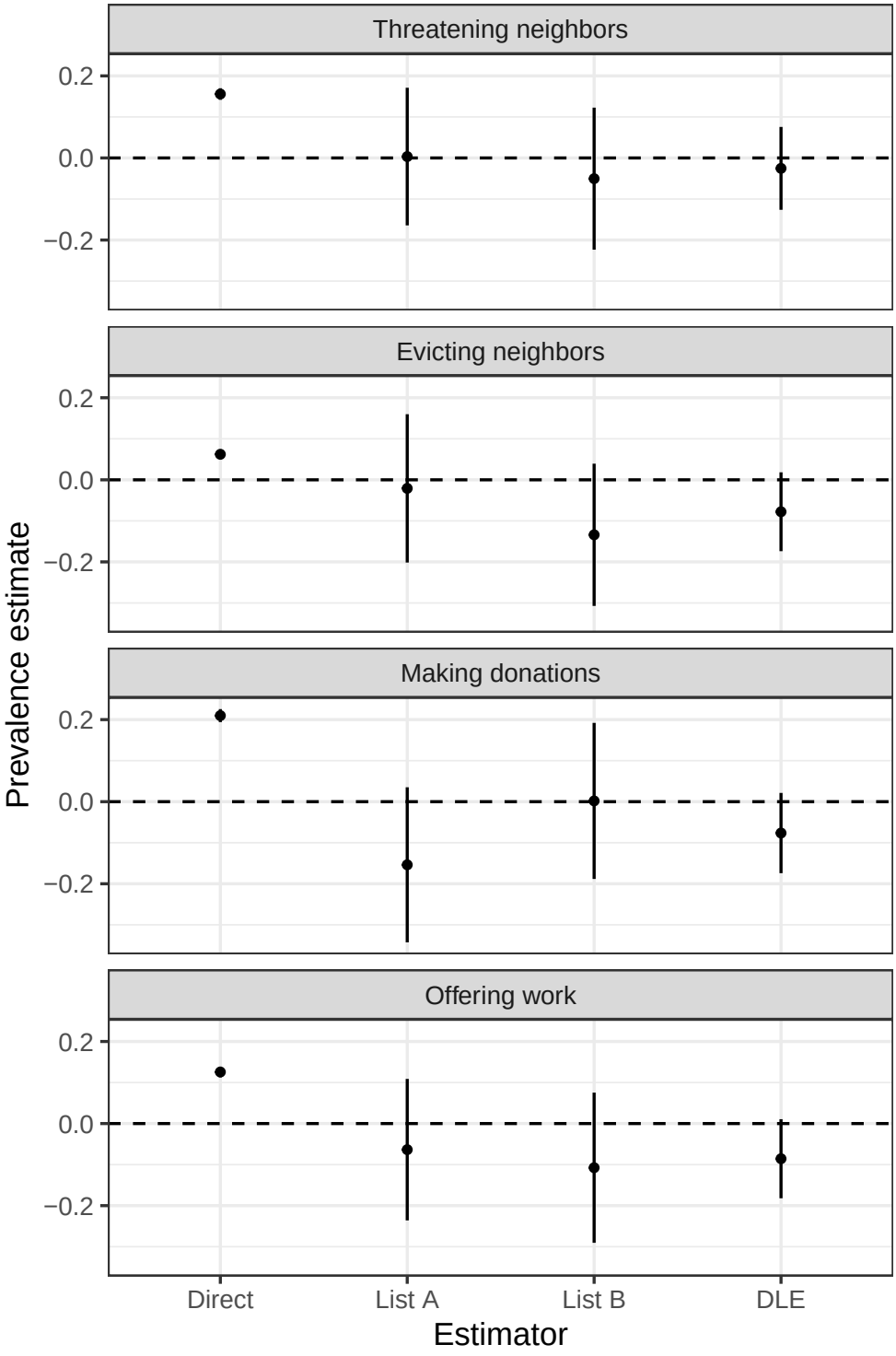

Figure 1: Prevalence estimates by sensitive item and estimator
Note: Vertical lines denote 95% confidence intervals.

Table 3: Statistical tests to rule out strategic response errors

Item	List	statistic	p.value	n
Placebo I: *Prevalence among "yes" in direct question*				
Donation	A	0.4615	0.048	133
Donation	B	-0.2588	0.272	133
Donation	Double	0.0837	0.548	133
Evict	A	0.4149	0.397	32
Evict	B	-0.1064	0.757	32
Evict	Double	0.1397	0.462	32
Offer Work	A	0.18	0.566	64
Offer Work	B	-0.1305	0.729	64
Offer Work	Double	0.0286	0.877	64
Threaten	A	0.1488	0.525	102
Threaten	B	0.3879	0.101	102
Threaten	Double	0.2447	0.075	102
Placebo II: *Effect of sensitive item placement on direct question*				
Donation	A	-0.032	0.323	635
Evict	A	0.0011	0.951	628
Offer Work	A	0.013	0.581	647
Threaten	A	0.0224	0.438	641
Carryover: *Effect of placing sensitive item in first list on second list responses*				
Donation	Double	-0.2357	0.155	668
Evict	Double	0.0847	0.576	664
Offer Work	Double	-0.1105	0.468	677
Threaten	Double	0.1159	0.407	671

3 Recovering Informative Estimates

If one can claim that the presence of a non-zero prevalence placebo is the primary source of the pattern of prevalence estimates in Figure 1, then the task is to reconstruct how responses would have looked like had the placebo item not been present. Because we did not include a direct question about the placebo item, we do not exactly which responses need to change.

This implies more than one valid prevalence estimate can be obtained by "cleaning" the responses contaminated by the placebo item. In other words, the estimand is no longer point-identified, but we can still compute bounds around it. As it is usually the case with non-parametric partial identification, sharp bounds are uninformative, but the list experiment setting provides additional structure that enables the computation of more informative bounds.

3.1 Assumptions

The goal in nonparametric partial identification is to identify the narrowest possible bounds using as little information as possible. In experiments, sharp bounds are derived exclusively by the observed data and standard experimental assumptions (Manski 1990). In this context, the non-interference part of SUTVA implies that only the responses in which the placebo item was included would be affected by its removal. One can compute lower and upper values of these sharp bounds by imputing the maximum (4 since that is the "updated" number of baseline items) and the minimum (0) value in the control responses, while leaving treated responses unchanged.[7]

The left column of Figure 2 shows the resulting sharp bounds computed in this manner, with the corresponding 95% confidence intervals based on 1,000 bootstrap replicates. As expected, the possible values far exceed the logical range of prevalence estimates, the lower

[7]The reversal is intentional. If only control responses are altered, then imputing the maximum value yields the smallest possible treatment effect, and viceversa.

bound even reaches up to a prevalence of about -2 across sensitive items.

How can one compute more informative estimates? The standard list experiment assumptions of no liars and no design effects provide additional structure on the possible values that responses would take in the absence of a placebo item. These assumptions are explicitly made about the sensitive item, but they are implicitly extended to the placebo item in the literature proposing placebo-controlled designs (**Agerberg2012?**; Riambau and Ostwald 2020). A zero-prevalence item only works as a placebo if one assumes that respondents will not actively select, which also means one also would expect responses to baseline lists to not be altered by the the inclusion of a placebo item in it.

Extending the standard list experimental assumptions to the placebo item has two implications when computing bounds: (1) Responses to control lists will only decrease (2) by a magnitude of one (unless the observed response is already zero). The lower bound in this context is calculated by decreased all observed control responses that already have a maximum value when the placebo is included (5). This is equivalent to making the conservative assumption of only decreasing by one the responses of those who unequivocally reveal holding the placebo item. The upper bound is then obtained by decreasing all observed control responses by one unit, which would be equivalent to assuming everyone holds the placebo item.

The right column in Figure 2 shows the bounds that follow from incorporating list experimental assumptions. These now approximate the logical range of a valid prevalence estimate, but are still uninformative in substantive terms.

3.2 Known placebo proportions

The bounds computed under the list experimental assumptions depict extreme cases of how control list responses would have been altered in the absence of a placebo item. The lower bound, derived by decreasing only the already maximum values by one unit, has the desirable

theoretical property of assuming only what one can observe in the data. However, the upper bound implies a placebo item that applies to every single unit, which is far from what a good placebo should be.

One way to impose a more realistic upper bound would be to restrict the proportion of units with lower than maximum control responses that are allowed to decrease their answer by one unit. If the researcher knew the proportion of known placebo in the target population from auxiliary information, they could impose a lower ceiling on the upper bound.

The new upper bound is then given by

$$\hat{\tau}_H = (1 - \delta) \times \bar{V}_{obs} + \delta \times \bar{V}_{max} \tag{1}$$

where δ is the proportion of known placebos, $\bar{V}_{obs}$ is the list experiment estimate based on observed data, and $\bar{V}_{max}$ is the list experiment estimate based on the vector of observations in which every control response is reduced by one unit. This equivalent to the standard approach to estimate partial identification bounds by computing a weighted average (Aronow and Miller 2019).

Figure 3 shows the resulting estimates with this new upper bound. Since we did not ask directly about the placebo item, nor we have access to reliable information from auxiliary data, we simulate estimates at discrete values in the $[0.05, 0.3]$ range. Overall, this produces much more informative upper bounds, especially at ranges in which one still consider the placebo item useful (e.g. 0.05, 0.1). However, the problem persists in the lower bound, as estimates still exceed the range of admissible values for a prevalence estimate.

Notice that if one were to ask direct questions about the placebo item, and be willing to claim it is measured with negligible error, then one would be able to exactly identify which responses should decrease by one unit, in which case the bounds would collapse into a point estimate again. In this case, the bounds proposed here may still be useful to assess the

sensitivity of estimates to potential measurement issues in the placebo item.

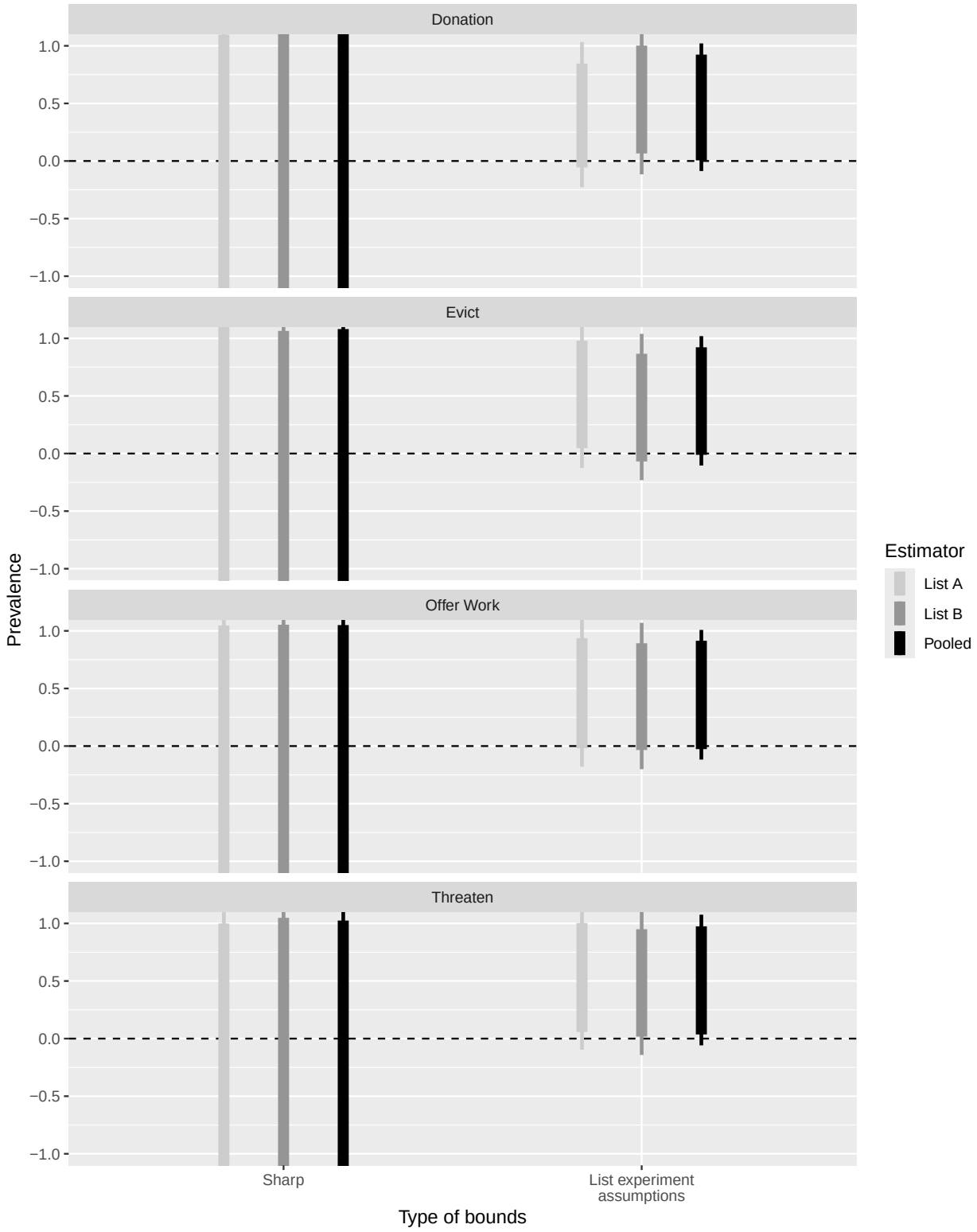

Figure 2: Nonparametric bounds by sensitive item
Note: Thinner lines indicate bootstrapped 95% confidence intervals. Zoomed in to facilitate visualization.

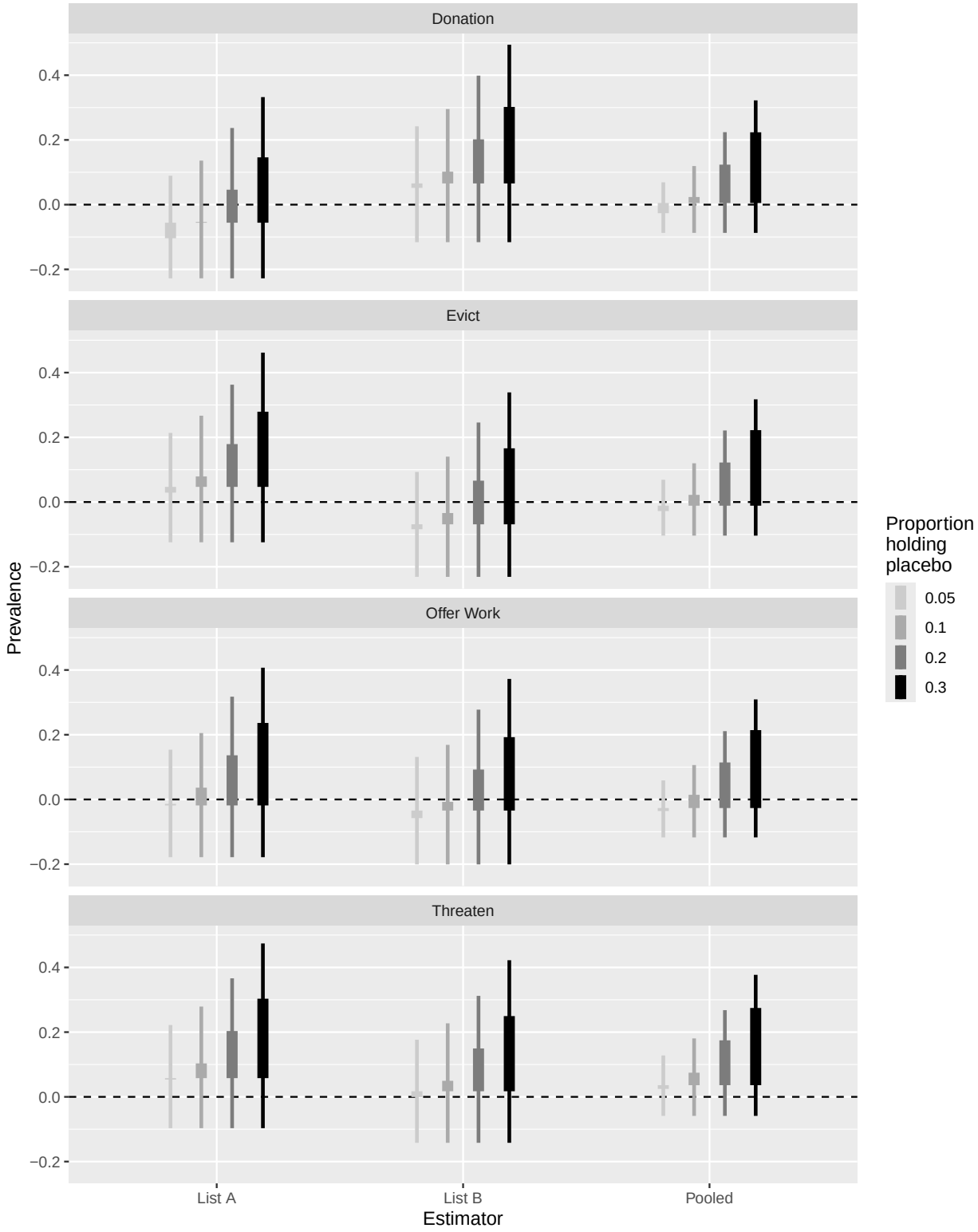

Figure 3: Nonparametric bounds by different assumed placebo proportions
Note: Thinner lines indicate bootstrapped 95% confidence intervals.

4 Conclusion

For now, here are a few thoughts about future steps:

1. The application of Placebo I feels shaky.
2. Any way to further refine the lower bounds? Our current thinking is that the Aronow et al. (2015) estimator that combines list with direct question estimates will narrow bounds further on both sides.
3. Our application is probably an extreme case, hence the bounds never look that good. Would it pay off to show the bounds in a more realistic application in which one has reason to suspect the placebo backfired?
4. Do the tests in Table 3 need to be adjusted for multiple testing?

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. 2017. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26 (1): 34–53.

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.

Arias, Enrique Desmond. 2017. *Criminal Enterprises and Governance in Latin America and the Caribbean*. Cambridge University Press. https://doi.org/10.1017/9781316650073.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3 (1): 43–66.

Aronow, Peter M., and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://doi.org/10.1017/9781316831762.

Auyero, Javier, and Katherine Sobering. 2019. *Ambivalent State: Police-Criminal Collusion at the Urban Margins*. Oxford University Press, Incorporated.

Barnes, Nicholas. 2021. "The Logic of Criminal Territorial Control: Military Intervention in Rio de Janeiro." *Comparative Political Studies* 55 (5): 789–831. https://doi.org/10.1177/00104140211036035.

Blair, Graeme, Winston Chou, and Kosuke Imai. 2019. "List Experiments with Measurement Error." *Political Analysis* 27 (4): 455–80.

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): 1297–1315.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77.

Blair, Graeme, Kosuke Imai, and Jason Lyall. 2014. "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan." *American Journal of Political Science* 58 (4): 1043–63.

Diaz, Gustavo. 2023. "Assessing the Validity of Prevalence Estimates in Double List Experiments." *Journal of Experimental Political Science*.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena 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*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, 185–210. New York: Wiley & Sons.

Flom, Hernán. 2022. *Informal Regulation of Criminal Markets in Latin America*. Cambridge University Press.

Fynn, Inés, Verónica Pérez Bentancur, and Lucía Tiscornia. 2024. "Uruguay 2023: Secu-

rity as a Persistent Challenge and the Decline of Non-Policy Politics as a Political Asset." *Revista de Ciencia Política (Santiago)*, no. ahead. https://doi.org/10.4067/s0718-090x2024005000108.

Gingerich, Daniel W., Virginia Oliveros, Ana Corbacho, and Mauricio Ruiz-Vega. 2016. "When to Protect? Using the Crosswise Model to Integrate Protected and Direct Responses in Surveys of Sensitive Behavior." *Political Analysis* 24 (2): 132–56. https://doi.org/10.1093/pan/mpv034.

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

Magaloni, Beatriz, Edgar Franco-Vivanco, and Vanessa Melo. 2020. "Killing in the Slums: Social Order, Criminal Governance, and Police Violence in Rio de Janeiro." *American Political Science Review* 114 (2): 552–72. https://doi.org/10.1017/s0003055419000856.

Manski, Charles. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80 (2): 319–23.

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.