Combining List Experiments and the Network Scale-Up Method to Improve Prevalence Estimates of Sensitive Attitudes and Behaviors

(Early draft, ask before citing)

Gustavo Díaz[*] Ines Fynn[†] Verónica Pérez Bentancur[‡]

Lucía Tiscornia[§]

April 2, 2025

Abstract

Social scientists use indirect questioning techniques like list experiments to estimate the prevalence of sensitive attitudes and behaviors through surveys. While list experiments reduce sensitivity bias compared to direct questions, they do so at the cost of increased variance. One proposal to alleviate this problem is combining the estimates of list experiments with direct questions to improve statistical precision. The problem is that in some applications researchers may be wary of including direct questions for practical or ethical reasons. In this case, we argue that questions from the Network scale-up method (NSUM) can be used in lieu of direct questions. Our paper illustrates how to combine single and double list experiments with the NSUM to improve statistical precision in the context of an application on the prevalence of criminal

[*]Assistant Professor of Instruction, Department of Political Science, Northwestern University. Email: gustavo.diaz@northwestern.edu

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

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

[§]Assistant Professor, School of Politics and International Relations, University College Dublin. Email: lucia.tiscornia@ucd.ie.

governance strategies in Montevideo, Uruguay. We also compare their performance with the combination of list experiments and direct questions.

1 Introduction

Social scientist use indirect questioning techniques in surveys to estimate the prevalence of sensitive attitudes and behavior in a population of interest. List experiments are a popular technique in political science, with topics including racial prejudice (Kuklinski et al. 1997), vote-buying (Gonzalez-Ocantos et al. 2011), and voter turnout (Holbrook and Krosnick 2010). While list experiments reduce sensitivity bias compared to direct questions, they do so at the cost of increased variance in prevalence estimates (Blair, Coppock, and Moor 2020; Rosenfeld, Imai, and Shapiro 2015). This means the downsides of implementing a list experiment may outweigh its advantages.

Several strategies exist to alleviate this problem and improve the position of list experiments along the bias-variance frontier. This paper focuses on the suggestion of combining list experiment estimates with auxiliary information to improve statistical precision. For example, Chou, Imai, and Rosenfeld (2017) combine indirect questioning techniques with aggregate-level information, Blair, Imai, and Lyall (2014) combine list experiments and endorsement experiments, and Aronow et al. (2015) combine list experiments with direct questions.[1]

Among the strategies that rely on auxiliary information to improve statistical precision, combining list experiments with direct questions is the most straightforward to implement, as it only requires one additional survey question. The logic is that, if the standard list experiment assumptions hold, one does not need to rely on a list experiment to learn about survey respondents who admit to the sensitive item when asked directly, which opens the way for a non-parametric weighted estimator of the prevalence rate (Aronow et al. 2015).

[1] Other strategies beyond the scope of this paper include inducing negative correlation among the baseline items in the list to minimize the possibility of ceiling or floor effects (Glynn 2013) or implementing a double list experiment (Miller 1984).

However, researchers may be wary of including direct questions in a survey for the same practical or ethical reasons that led them to consider a list experiment to begin with. In practice, if the researcher anticipates a question to be sensitive enough to warrant indirect questioning, then one may expect a relative low proportion of positive responses to the direct question, which would lead to an uninformative combined estimator. On ethical grounds, researchers may be wary of including direct questions around topics that may elicit negative emotions or force participants to relive trauma (Fujii 2012).

In this case, we argue that questions from the Network scale-up method (NSUM) can be used in lieu of direct questions. The NSUM estimates the size of hard to reach populations by asking respondents how many people they know who hold a sensitive behavior of interest. By calibrating the size of an individual's network with anchor questions (e.g. "How many people named Silvia do you know"), one can claim that individuals with higher than average exposure to the sensitive trait relative the size of their network are exposed to sensitive trait themselves (Laga, Bao, and Niu 2021).

Using the NSUM to improve list experiment estimates has two advantages. First, this approach is less sensitive than direct questions, so one can implement it without fear of imposing an undue burden on study participants. Second, unlike other indirect questioning techniques, the NSUM allows researchers to identify those who openly admit to the sensitive item, which allows the implementation of a non-parametric estimator of the prevalence rate, whereas other indirect questioning techniques would need additional modeling assumptions to combine with list experiments.

Our paper illustrates how to combine single and double list experiments with the NSUM to improve statistical precision in the context of an application on the prevalence of criminal governance tools in Uruguay. We also compare their performance with the combination of list experiments and direct questions. Our work contributes to a growing literature seeking to improve the efficiency of list experiments (Glynn 2013; Blair, Imai, and Lyall 2014; Aronow et al. 2015; Chou, Imai, and Rosenfeld 2017; Diaz 2023). This improves researchers' ability to detect and estimate sensitive attitudes and behaviors with surveys and facilitates the application of the technique to a broader range of settings.

2 Indirect questioning techniques

2.1 List experiment with direct question

List experiments can reduce misreporting on sensitive questions in surveys by asking about the sensitive attitude or behavior or interest indirectly. As a running example, consider part of an original list experiment conducted in Montevideo, Uruguay [self-cite]. In the canonical design, after assignment to conditions, respondents in the control group are asked the following question:

`Please tell us how many of these things happened to you in your neighborhood in the last six months. We just want to know how many things, not which ones.`

- `I saw people doing sports`
- `I visited friends`
- `I participated in activities organized by feminist groups`
- `I went to church`

The treatment group is asked the same question, but the baseline list now includes one additional item:

- `I saw gang members threatening neighbors`

Which is the sensitive attitude or behavior of interest. With this design, one can estimate the difference-in-means

$$\widehat{V} = E[V_i(1)|Z_i = 1] - E[V_i(0)|Z_i = 0] \tag{1}$$

where $\widehat{V}$ is interpreted as the proportion of respondents holding the sensitive item. $V_i(*)$ denotes observed responses to the list experiment question with (1) and without (0) the sensitive item, and Z_i denotes whether the respondent was assigned to the treatment (1) or control (0) list.[2]

$\widehat{V}$ is a valid estimator of the prevalence of the sensitive item under the standard experimental assumptions, plus two additional assumptions. First, the "no liars" assumption requires that respondents do not lie about holding the sensitive item when they do not. Second, the "no design effects" assumption states that latent responses to the baseline list do not change when the sensitive item is included (Blair and Imai 2012). Previous work has proposed tools to diagnose potential violations to these assumptions and adjust estimates accordingly (Blair and Imai 2012; Li 2019; Diaz 2023).

The advantage of implementing a list experiment is a reduction in sensitivity bias, which translates to point estimates closer to the true prevalence of the sensitive item in the population of interest. While this bias reduction is usually unobservable, a validation study suggests this is the case (Rosenfeld, Imai, and Shapiro 2015). The downside of implementing list experiments is a high-variance estimator, which translates to wider confidence intervals and lower statistical power relative to direct questioning. The implication of this bias-variance tradeoff is that, in many social science applications, list experiments produce estimates that are statistically indistinguishable from those produced by direct questioning (Blair, Coppock, and Moor 2020).[3]

[2] This notation is a simpler version from the notation in Aronow et al. (2015). See this piece and Blair and Imai (2012) for more complete treatments on the design and analysis of list experiments.

[3] This tradeoff also applies to other indirect questioning techniques that reduce sensitivity bias by intro-

Previous work has proposed several strategies to address the low statistical precision of the standard list experiment estimator. These techniques involve complementing list experiments with additional information. For example, Chou, Imai, and Rosenfeld (2017) show how to incorporate population-level information (e.g. election results by county) to reduce variance in the individual-level estimation of the sensitive trait of interest, and Blair, Imai, and Lyall (2014) show how to combine estimates from list experiments and endorsement experiments.

More relevant to this paper, Aronow et al. (2015) show how to combine list experiments estimates with direct questions. Let $Y_i \in \{0, 1\}$ denote individual responses to the direct question. In our running example, the survey asked:

`Over the last six months, have you seen criminal gangs threatening neighbors? (Yes/No)`

Under the direct question, the expected proportion of positive responses $\widehat{Y} = E[Y_i]$ is a straightforward estimator of the sensitive item prevalence in the population of interest.

The core idea in Aronow et al. (2015) is that we do not need to use indirect question responses for those who openly admit to the sensitive item. This informs the combined prevalence estimator

$$\hat{\mu} = \overline{Y} + (1 - \overline{Y})(\overline{V}_{Y_i=0}) \tag{2}$$

where $\overline{Y}$ and $\overline{V}$ are the sample analogues of $\widehat{Y}$ and $\widehat{V}$, respectively. This estimator is equivalent to a weighted average of the prevalence rates calculated with the two methods, using the responses to the list experiment question only among those who did not openly admit to the sensitive item in the direct question, $Y_i = 0$ (see Aronow et al. 2015 for details).

This approach is appealing because it does not rely on administrative or contextual information that may not be available in every application. Across empirical applications, Aronow et

ducing noise in the estimation (Blair 2015).

al. (2015) document reductions in sampling variance ranging from 14 to 67% relative to only asking list experiment questions. Another advantage is that researchers can always combine direct questions with estimates drawn from multiple indirect questioning techniques, such as list with endorsement experiments (Blair 2015) or, as we show below, with double list experiments (Glynn 2013; Diaz 2023).

The main drawback of this approach is that researchers may have reservations about including direct questions in the survey to begin with. This may occur due to practical or ethical considerations. In practice, if the researcher anticipates a question to be sensitive enough to warrant indirect questioning, then one may expect a relative low proportion of positive responses to the direct question, which would lead to an uninformative combined estimator. On ethical grounds, researchers may be wary of including direct questions around topics that may elicit negative emotions or force participants to relive trauma (Fujii 2012).

To implement this approach in contexts where direct questions are not admissible, one needs an indirect questioning technique that allows researchers to infer individual-level responses to the sensitive item for every respondent while still guaranteeing anonymity. This is challenging since most indirect questioning techniques rely on masking individual responses. The next section outlines and indirect questioning technique that allow researchers to accomplish this goal with minimal assumptions.

2.2 Network scale-up method as a replacement for direct questions

The network scale-up method (NSUM) is a technique used to estimate the size of hard-to-reach populations in surveys. The original motivation behind this technique was to estimate the number of casualties after an earthquake (Bernard et al. 1991). This is challenging because individuals experience different levels of exposure to the event of interest based on their location and the density of their personal network. Moreover, the population of interest is usually concentrated in a manner unknown to the researcher, which makes representative sampling impractical or uninformative.

While the original motivation did not consider sensitive topics, the NSUM was quickly adopted in the health sciences to estimate the size of populations that remain hidden due to misreporting and the absence of a sampling frame, such as HIV positive patients, sex workers, and illegal drug users (e.g. Guo et al. 2013; Jing et al. 2018; Salganik et al. 2011).

The NSUM uses numerous "How many X do you know?" questions to create aggregated relational data at the respondent level. These data are then used to estimate network features without observing networks directly, such as the size each respondent's personal network or the size of a hidden population of interest within a network.

For example, the survey in Uruguay asked:

`How many people do you know, who also know you, with whom you have interacted in the last year in person, by phone, or any other channel?`

- `Who are public employees`
- `Who are registered members of a political party`
- `Who have children attending public school`
- `Who are in prison`

⋮

- *`Who have been threatened by gang members`*

Appendix XX shows the complete list of 15 baseline items. For each item, respondents choose an integer ranging from zero to ten or more. Responses are truncated at the upper limit to avoid placing excessive weight on individuals with a high number of nodes in the network (Zheng, Salganik, and Gelman 2006).

The main feature of NSUM is that it allows researchers to draw population-level estimates of size of a subpopulation of interest with a convenience sample (see Laga, Bao, and Niu 2021; McCormick 2020 for reviews; and Killworth et al. 1990; Zheng, Salganik, and Gelman 2006 for technical details). However, this relies on the assumption that the ratio of the subpopulation of interest to the general population is equivalent to the proportion of an individual's network who belongs to the key subpopulation

$$\frac{N_k}{N} = \frac{y_{ik}}{d_i} \tag{3}$$

where N is the population size, N_k is the true size of subpopulation k, y_{ik} is individual i's response to the "how many" question for group k, and d_i is the overall degree or number of edges for individual i. This is known as the constant proportion assumption. In most applications, all quantities are unknown except for y_{ik}.

If the constant proportion assumption is violated, then NSUM estimates are biased relative to the true population parameter of interest. The most common violation of constant proportion are barrier effects, which occur when respondents are more or less likely to know someone across different subpopulations due to their own characteristics (Laga, Bao, and Niu 2021). For example, people are more (less) likely to know someone who has been threatened by gang members in high (low) crime victimization areas.

Because barrier effects are likely in social science applications, political scientists have used the NSUM mostly to reconstruct individual-level network features without observing networks directly. For example, Calvo and Murillo (2012) use the NSUM to calibrate the number of registered members of a political party that a person knows. Ventura, Ley, and Cantú (2023) use a similar approach to identify individual levels of exposure to crime victimization. In both cases, this information is eventually used as covariates in regression models.

We follow the same procedure to identify individuals who can be considered as holding a sensitive trait of interest. The goal is to identify individuals with unusually high exposure to the sensitive network of interest relative to the size of their personal network, we do so with the hierarchical model of observed responses to the "how many" questions proposed by Zheng, Salganik, and Gelman (2006)

$$y_{ik} \sim \text{negative-binomial}(e^{\alpha_i+\beta_k}, \omega_k) \tag{4}$$

where α_i is the (logged) total degree of individual i across all groups, β_k is the (logged) total number of links that involve group k across individuals, and ω_k is an overdispersion parameter that captures the variance in the probability of knowing someone in group k across individuals.[4] The main advantage of this approach is that it models overdispersion directly instead of considering it a violation of the constant proportion assumption.

Previous applications in political science have used this model to estimate key parameters and then predict their values at the individual level. For example, Ventura, Ley, and Cantú (2023) uses a two-step maximum-likelihood procedure to compute standardized residuals on for the "how many" questions across individuals and groups.

$$r_{ik} = \sqrt{y_{ik}} - \sqrt{e\alpha_i + \beta_k} \tag{5}$$

The first step uses groups that the researchers expect to have high recall among respondents to estimate α_i with high precision. This value is included in the second step along with the questions about the remaining groups, which includes those the researcher is interested on.[5]

The residual r_{ik} can be interpreted as the distance between the number of people in group k known by individual i, relative to the average size of group k, given the size of individual's

[4]This is an unconventional parameterization of the traditional negative binomial function $y \sim \text{NB}(r, p)$ to make the overdispersion parameter more visible. See Zheng, Salganik, and Gelman (2006) for further justification.

[5]The model could also be fitted in one step, but the NSUM setup lends itself naturally to the distinction between groups of interest and those used for calibration.

i personal network. In other words, higher values of r_{ik} indicate that individual i knows a disproportionately higher number of people in group k.

Since the residual r_{ik} is a continuous measure, we need a rule to convert it into a binary indicator of whether individual i experiences unusually high exposure to group k, relative to their personal network and overall size of the group

$$Y_i' = \begin{cases} 1, & \text{if } r_{ik} > \overline{r}_{ik} + \lambda S_{r_{ik}} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where Y_i' is an alternative measure of Y_i calculated with NSUM instead of the direct question, the bar operator denoting the mean, and S denoting the standard deviation. The tuning parameter λ governs what is considered unusually high exposure to the group of interest. For example, with $\lambda = 0$, everyone with an above-average residual would be considered to have unusually high exposure. As a safeguard, we suggest $\lambda = 1$ as a conservative rule of thumb, which we report in our application.

Note that λ implies a bias-variance tradeoff. Larger values imply one can more credibly justify that units with $Y_i' = 1$ can be considered as facing unusually high exposure. However, they also imply fewer observations, we would make the weighted average estimator less informative. In practice, we recommend researchers perform sensitivity analysis to quantify the robustness of their chosen value for λ (see Appendix XX for an example).

After computing Y_i', one can simply use it instead of the direct question indicator Y_i to estimate $\widehat{Y}$ and, ultimately, the combined estimator $\hat{\mu}$.

However, to use the exposure indicator Y_i' in this way, one needs to assume that units with unusually high exposure to a group of people holding a sensitive trait of interest are also likely to hold the sensitive trait themselves. We call this assumption **symmetrical exposure** [accepting suggestions for a fancier name]. This assumption is plausible for sensitive questions that relate to individual experiences (e.g. witnessing criminal activity), but may not necessarily apply to personal beliefs or attitudes (e.g. being opposed to certain groups moving into one's neighborhood).

[TO DO: Write assumption in math]

As an assumption, symmetrical exposure cannot be tested or confirmed, but careful question wording may help the researcher justify it. For example, the list experiment and direct questions in the Uruguay survey are worded as *seeing* criminal activity in the neighborhood, while the NSUM question is worded as *knowing someone* who has experienced it. Both approaches to phrasing tap into the same construct of interest, which is exposure to highly localized criminal governance tools.

3 Application: Criminal Governance in Uruguay

3.1 Survey

[Self-cite] conducted an online survey on a sample of 2,688 residents in the city of Montevideo, Uruguay in 2022. The main substantive goal was to document the prevalence of criminal governance tools in a context of high state capacity, in which one would expect lower crime rates relative to other countries in the region, but criminal governance still persists on pockets of the city (see Barnes 2017; Lessing 2021 for overviews on criminal governance).

The survey includes a double list experiment (DLE) on four different sensitive items. Respondents are presented with two baseline lists as separate questions in the survey flow. In each, respondents are asked how many items, but not which ones, apply to their situation. Table 1 shows the two baseline lists.

Respondents were randomly assigned to see one of four sensitive items, each depicting either a negative or positive criminal governance tool:

1. Saw gang members threatening neighbors

Table 1: Double list experiment baseline lists in Uruguay survey

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

2. Saw gang members evicting neighbors from their homes
3. Saw gang members making donations to neighbors
4. Saw gang members offering work to neighbors

The sensitive item then appears either in list A or list B, which is the standard double list experiment set up (Miller 1984).[6] This enables three list-experiment prevalence estimators:

1. Difference-in-means using responses to list A only
2. Difference-in-means using responses to list B only
3. Average of the differences-in-means computed with list A and list B (DLE)

That means we have $4 \times 3 = 12$ list-experiment prevalence estimates. Each sensitive item was also asked as a direct question of every respondent, even if they do not see in the list experiment questions, and as part of a NSUM question, where they only see the same sensitive item that appears in the list experiment. That means we have two additional ways to compute a prevalence estimate for each combination of sensitive items and list-experiment estimator, since each can be combined with the corresponding direct question or NSUM question. That expands to $4 \times 3 \times 3 = 36$ different points of comparison to assess the performance of NSUM questions.

[6] Following recent suggestions in the literature (Agerberg and Tannenberg 2021; Riambau and Ostwald 2020), the list that does not include the sensitive item included a placebo item that read "I did not drink mate." This should have been a near-zero prevalence item that would help adjust for measurement error attributable to different list lengths. However, this inadvertently introduced attenuation bias in in list-experiment estimates. We address this issue elsewhere.

3.2 Combining estimates

Table 2 shows the results of every estimation procedure in the rows for each sensitive item in the columns. Each point estimate can be interpreted as the proportion of individuals who have witnessed criminal organizations engaging in each of the listed activities in the last six months. Numbers in brackets represent 95% confidence intervals, computed by analytic derivation for every case, except for the combination of DLE + NSUM questions, for which the analytic derivation has not been documented yet. In this cases, confidence intervals are computed with bootstrapped standard errors based on 5,000 resamples per estimate.

Starting with the single-technique estimates in Table 2, all estimates based on the list experiment question exhibit the attenuation bias referenced in footnote 6 due to the inclusion of a placebo sensitive item. This is a separate issue that we address elsewhere. The other single-technique estimates (direct, NSUM) perform as one would expect, with prevalence estimates ranging from 6.2% (direct, evict) to 27.1% (NSUM, work). As a general pattern, NSUM prevalence estimates tend to be higher than the estimates using the direct question, with the exception of having seen gangs threatening neighbors, for which the estimates are similar. This occurs despite using a somewhat conservative approach to determine who is considered as exposed to each sensitive item under the NSUM questions.

Moving on to the estimates that combined list experiments estimates with direct or NSUM questions, we find broadly the same pattern, combined estimators relying on NSUM questions tends to yield higher prevalence estimates than their direct question counterparts. Once again, the exception are the estimates for having witnesses gangs threatening neighbors, in which case combined estimates based on direct questions are higher.

Taken together, these results illustrates that NSUM questions can be used in lieu of direct questions to produce improve the estimation of list experiments. However, they are not a direct replacement, as they yield different estimates. Moreover, without a concrete behavioral benchmark, we cannot determine whether they yield a closer approximation to the ground

Table 2: Prevalence estimates by sensitive item and technique

	Threaten	Evict	Donate	Work
Single-technique estimates				
List A	−0.003	−0.022	−0.151	−0.064
	[−0.171, 0.165]	[−0.203, 0.159]	[−0.340, 0.038]	[−0.237, 0.108]
List B	−0.048	−0.134	−0.002	−0.107
	[−0.221, 0.125]	[−0.307, 0.039]	[−0.192, 0.189]	[−0.290, 0.076]
DLE	−0.025	−0.078	−0.076	−0.086
	[−0.126, 0.076]	[−0.174, 0.018]	[−0.174, 0.021]	[−0.182, 0.011]
Direct	0.156	0.062	0.210	0.126
	[0.142, 0.170]	[0.053, 0.072]	[0.194, 0.226]	[0.113, 0.138]
NSUM	0.149	0.207	0.271	0.271
	[0.117, 0.181]	[0.169, 0.244]	[0.230, 0.311]	[0.230, 0.312]
Combined with direct question				
List A + direct	0.158	0.032	0.091	0.043
	[0.004, 0.312]	[−0.136, 0.200]	[−0.072, 0.255]	[−0.120, 0.205]
List B + direct	0.120	−0.074	0.209	0.004
	[−0.037, 0.278]	[−0.243, 0.094]	[0.043, 0.375]	[−0.168, 0.175]
DLE + direct	0.134	−0.011	0.149	0.051
	[0.047, 0.221]	[−0.100, 0.079]	[0.072, 0.227]	[−0.033, 0.135]
Combined with NSUM questions				
List A + NSUM	0.147	0.189	0.161	0.224
	[0.004, 0.289]	[0.050, 0.328]	[0.013, 0.308]	[0.088, 0.360]
List B + NSUM	0.109	0.100	0.270	0.193
	[−0.044, 0.261]	[−0.042, 0.243]	[0.121, 0.418]	[0.052, 0.333]
DLE + NSUM	0.128	0.145	0.215	0.208
	[0.036, 0.219]	[0.058, 0.231]	[0.127, 0.303]	[0.122, 0.295]

95% confidence intervals in brackets

truth.

The original goal of combining list experiments with direct questions is to increase the statistical precision of the list experiment estimator (Aronow et al. 2015). Therefore, another metric to evaluate is whether using the NSUM as a replacement for direct questions yields comparable levels of variance reduction.

Table 3 shows descriptive evidence in this regard. Each cell represents the ratio of variances between the combined estimator and the corresponding single list estimator (e.g. the first cell corresponds to $\frac{Var(\text{List A + direct})}{Var(\text{List A})}$). Values below one indicate that the combined estimator has lower variance than the corresponding single-technique estimator, which translates to smaller standard errors or narrower confidence intervals. In general terms, both direct and NSUM questions achieve comparable degrees of variance reduction, with NSUM performing slightly better for single-list estimates, and direct questions performing slightly better for DLE estimates, although we do not have enough evidence to determine if this pattern holds beyond the current application.[7]

[7] TO DO: More formal tests to determine whether differences in variance reduction are significant.

Table 3: Variance reduction relative to single-technique in prevalence estimates by technique and sensitive item

	Combined with direct question			Combined with NSUM		
Sensitive item	List A	List B	DLE	List A	List B	DLE
Threaten	0.915	0.910	0.862	0.846	0.885	0.907
Evict	0.929	0.976	0.935	0.769	0.824	0.903
Donate	0.868	0.873	0.794	0.785	0.779	0.898
Work	0.946	0.938	0.872	0.791	0.769	0.898
Average	**0.915**	**0.924**	**0.866**	**0.798**	**0.814**	**0.902**

4 Conclusion

Coming soon!

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.

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.

Barnes, Nicholas. 2017. "Criminal Politics: An Integrated Approach to the Study of Organized Crime, Politics, and Violence." *Perspectives on Politics* 15 (4): 967–87. https://doi.org/10.1017/s1537592717002110.

Bernard, H. Russell, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. 1991. "Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results." *Social Science Research* 20 (2): 109–21. https://doi.org/10.1016/0049-089x(91)90012-r.

Blair, Graeme. 2015. "Survey Methods for Sensitive Topics." *Comparative Politics Newsletter* 12: 44.

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.

Calvo, Ernesto, and Maria Victoria Murillo. 2012. "When Parties Meet Voters." *Comparative Political Studies* 46 (7): 851–82. https://doi.org/10.1177/0010414012463882.

Chou, Winston, Kosuke Imai, and Bryn Rosenfeld. 2017. "Sensitive Survey Questions with Auxiliary Information." *Sociological Methods & Research* 49 (2): 418–54.

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

Fujii, Lee Ann. 2012. "Research Ethics 101: Dilemmas and Responsibilities." *PS: Political Science &Amp; Politics* 45 (4): 717–23. https://doi.org/10.1017/s1049096512000819.

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

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Meléndez, Javier Osorio, and David W. Nickerson. 2011. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56 (1): 202–17.

Guo, Wei, Shuilian Bao, Wen Lin, Guohui Wu, Wei Zhang, Wolfgang Hladik, Abu Abdul-Quader, Marc Bulterys, Serena Fuller, and Lu Wang. 2013. "Estimating the Size of HIV Key Affected Populations in Chongqing, China, Using the Network Scale-up Method." Edited by D. William Cameron. *PLoS ONE* 8 (8): e71796. https://doi.org/10.1371/journal.pone.0071796.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74 (1): 37–67.

Jing, L., Q. Lu, Y. Cui, H. Yu, and T. Wang. 2018. "Combining the Randomized Response Technique and the Network Scale-up Method to Estimate the Female Sex Worker Population Size: An Exploratory Study." *Public Health* 160 (July): 81–86. https://doi.org/10.1016/j.puhe.2018.04.002.

Killworth, Peter D., Eugene C. Johnsen, H.Russell Bernard, Gene Ann Shelley, and Christopher McCarty. 1990. "Estimating the Size of Personal Networks." *Social Networks* 12 (4): 289–312. https://doi.org/10.1016/0378-8733(90)90012-x.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41 (2): 402.

Laga, Ian, Le Bao, and Xiaoyue Niu. 2021. "Thirty Years of the Network Scale-up Method." *Journal of the American Statistical Association* 116 (535): 1548–59. https://doi.org/10.1080/01621459.2021.1935267.

Lessing, Benjamin. 2021. "Conceptualizing Criminal Governance." *Perspectives on Politics* 19 (3): 854–73. https://doi.org/10.1017/s1537592720001243.

Li, Yimeng. 2019. "Relaxing the No Liars Assumption in List Experiment Analyses." *Political Analysis* 27 (4): 540–55.

McCormick, Tyler H. 2020. "The Network Scale-up Method." Edited by Ryan Light and James Moody. *The Oxford Handbook of Social Networks*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190251765.013.14.

Miller, Judith Droitcour. 1984. "A New Survey Technique for Studying Deviant Behavior." PhD thesis, George Washington University.

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.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2015. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60 (3): 783–802.

Salganik, Matthew J., Dimitri Fazito, Neilane Bertoni, Alexandre H. Abdo, Maeve B. Mello, and Francisco I. Bastos. 2011. "Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence from a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil." *American Journal of Epidemiology* 174 (10): 1190–96. https://doi.org/10.1093/aje/kwr246.

Ventura, Tiago, Sandra Ley, and Francisco Cantú. 2023. "Voting for Law and Order: Evidence from a Survey Experiment in Mexico." *Comparative Political Studies*, May, 001041402311690. https://doi.org/10.1177/00104140231169035.

Zheng, Tian, Matthew J Salganik, and Andrew Gelman. 2006. "How Many People Do You Know in Prison?" *Journal of the American Statistical Association* 101 (474): 409–23. https://doi.org/10.1198/016214505000001168.