Assessing the Validity of Prevalence Estimates in Double List Experiments*

Gustavo Diaz†

February 13, 2022

Link to most recent version

Abstract

Social scientists use list experiments in surveys to estimate the prevalence of sensitive attitudes and behaviors in a population of interest. However, the cumulative evidence suggests that list experiment estimator is underpowered to capture the extent of sensitivity bias in common applications. The literature suggests double list experiments (DLEs) as an alternative to improve along the bias-variance frontier. This variant of the research design brings the additional burden of justifying the list experiment identification assumptions in both lists, which raises concerns over the validity of DLE estimates. To overcome this difficulty, this paper outlines two statistical tests to detect unintended respondent behavior that follows from violations to the identification assumptions. I illustrate their implementation with data from a study on support toward anti-immigration organizations in California and explore their properties via simulation.

*I thank Jake Bowers, Alex Coppock, Luke Sanford, CIPR workshop participants, and audiences at the 2020 NYU CESS Conference on Experimental Political Science, Polmeth 2021, APSA 2021, and Latin American Polmeth 2021 for valuable feedback.

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

1 Introduction

Social scientists use list experiments in surveys to estimate the prevalence of sensitive attitudes and behaviors in a population of interest, with topics including racial prejudice (Kuklinski, Cobb, and Gilens 1997), vote-buying (Gonzalez-Ocantos et al. 2011), sexual behavior (Chuang et al. 2021), and voter turnout (Holbrook and Krosnick 2010). A recent review shows that the standard difference in means estimator in the list experiment is underpowered to capture the extent of sensitivity bias in common applications. This happens because the bias reduction of list experiments relative to direct questioning comes at the cost of increased variance (Blair, Coppock, and Moor 2020).

The literature proposes double list experiments (DLEs) as an alternative research design to improve along the bias-variance frontier. DLEs consist of two parallel list experiments implemented simultaneously, with the average of the treatment effects in each experiment as an estimator of the prevalence of the sensitive trait. Because in this design every respondent sees the sensitive item once, the variance of the pooled DLE estimate is, in expectation, reduced by half (Droitcour et al. 1991).

While DLEs promise more precise estimates by just adding an extra survey question, they are yet to become widespread practice. This is because they bring the additional burden of justifying the list experiment assumptions for two lists of baseline items. This is a challenge considering research that shows how using different baseline lists often yields diverging prevalence estimates of the same sensitive behavior (Chuang et al. 2021).

This paper outlines two statistical tests to detect whether participants alter their responses in ways that would violate the identification assumptions in a DLE. For example, respondents may deflate (inflate) their responses to baseline items to avoid association with a socially undesirable (desirable) sensitive item (Zigerell 2011).

Both tests leverage variation in the timing with which the sensitive item is presented to

respondents in DLEs. I refer to these as treatment schedules. In a DLE, respondents see the two baseline lists and the sensitive item appears at random in the first or second list. When respondents see the sensitive item in the first list, they can alter their response to both lists. When they see the sensitive item in the second list, they can only alter their response to that list. By comparing the association between responses across treatment schedules, one can detect unintended response patterns, which helps in assessing the validity of prevalence estimates.

More formally, I propose the difference in differences and Stephenson's signed rank (Stephenson 1981) as tests to detect whether one can attribute extreme responses to the variation in treatment schedules. I illustrate the application of these tests with a reanalysis of a DLE on support for anti-immigration organizations in California (Alvarez et al. 2019) and examine their properties via simulation.

2 Double list experiments: Promise and challenge

As a running example, consider the study by Alvarez et al. (2019) on support for anti-immigration organizations in California.[1] Participants in an online survey in 2014 were asked to indicate how many of the following organizations they support (without specifying which ones):

- `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)`

[1] This study also appears in Li (2019). For a formal treatment of list experiments and DLEs see Blair and Imai (2012), Droitcour et al. (1991), and Glynn (2013).

- `American Red Cross (humanitarian organization)`

In the standard list experiment, the control group sees the list as it appears above. The treatment group also sees the following sensitive item:

- `Organization X (organization advocating for immigration reduction and measures against undocumented immigration)`

Respondents saw the name of real organizations, but the replication materials censor them for ethical reasons. In the standard list experiment, the difference in means between treatment and control estimates the proportion of the target population who supports Organization X. This estimator is valid under standard experimental assumptions, plus two more (Blair and Imai 2012). First, respondents do not misreport holding the sensitive trait (no liars). This assumption is violated if respondents who hold the sensitive trait give exactly the same response under treatment and control. Li (2019) develops estimate bounds that allow researchers to relax this assumption.

Second, participants do not alter their response to the baseline items when the sensitive item is included (no design effects), which is violated when respondents deflate (inflate) their responses to avoid (emphasize) association with the sensitive item (Zigerell 2011). Blair and Imai (2012) propose a test to detect potential violations of this assumption in the standard list experiment.[2]

A recent meta-analysis shows that the list experiment estimator is underpowered to detect sensitivity biases in common applications (Blair, Coppock, and Moor 2020). This is because of the bias-variance trade-off. A validation study shows that, compared to direct questioning, list experiments produce estimates closer to the true prevalence, albeit with wider confidence intervals (Rosenfeld, Imai, and Shapiro 2015).

[2]Aronow et al. (2015) characterize both no liars and no design effects as a single monotonicity assumption, under which individual potential outcomes under treatment are never smaller than potential outcomes under control.

An underexplored solution to reduce variability in estimates without compromising bias reduction is to implement a DLE (Droitcour et al. 1991). A DLE differs from the standard list experiment in two ways. First, DLEs include two lists of baseline items as separate questions, usually close to each other in the survey flow.

Continuing with the running example, Alvarez et al. (2019) include a second list:

- `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)`

For simplicity, these are list A and B. The second way in which DLEs differ from the standard design is that the sensitive item is randomly assigned to appear in A or B. This is equivalent to conducting two parallel list experiments. Some respondents receive A under treatment and B under control, others receive A under control and B under treatment.

This implies one difference in means for each list. The DLE estimator is the average of these two. Because each respondent serves as both treatment and control in parallel experiments, DLE estimates have roughly half of the variance of the single-list estimator (Droitcour et al. 1991).

DLEs promise increased precision at the cost of one additional survey question and no additional assumptions. However, one must now justify these assumptions for two lists. The challenge is that single-list prevalence estimates can vary considerably across comparable lists (Chuang et al. 2021). Since single-list estimates often have wide confidence intervals, a likely scenario is to find different point estimates with confidence intervals that overlap. This means one cannot determine whether the average of the two is a credible approximation of the true prevalence.

The design in Alvarez et al. (2019) helps to illustrate this point. The study also includes a second sensitive item:

- `Organization Y (citizen border patrol group combating undocumented immigration)`

Organizations X and Y are mutually exclusive, so one can analyze them as separate DLEs. Since respondents always see list A first, the experiment has four possible combinations of sensitive items and their placement, these appear in Table 1. Each experiment has, three different estimates: Two single-list estimates and the pooled DLE estimate.

Figure 1 shows these estimates for both sensitive items. For Organization X, all estimators suggest a non-zero prevalence rate around 0.3. For organization Y, estimates vary more. The estimate for list A suggest a prevalence of 0.1 that is indistinguishable from zero, list B suggests a non-zero prevalence of 0.4, and the pooled DLE estimate suggest a non-zero prevalence of 0.3.

The baseline lists do not change across organizations, so the differences in estimates may come from unintended responses to the sensitive item. Organization Y is a group attempting to take matters against undocumented immigration into their own hands rather than just pushing for stricter policies, so it is more likely to stand out to respondents.

Since list A always appears first in this study, the pattern of estimates for Organization Y suggests response deflation. In the list experiment for list A, only the treatment group sees the sensitive item, so they deflate their responses to avoid signaling support for Organization Y. Since the control group has not seen the sensitive item yet, they respond truthfully, which biases the single-list estimate toward zero. In contrast, since list B always appears second, both treatment and control group have already seen the sensitive item, so both shift in the same direction.

Since the confidence intervals for the single-list estimates for list A and B in the Organization

Table 1: Research Design in Alvarez et al (2019)

	Placement	
	List A	List B
Sensitive item		
Organization X	545	525
Organization Y	537	543

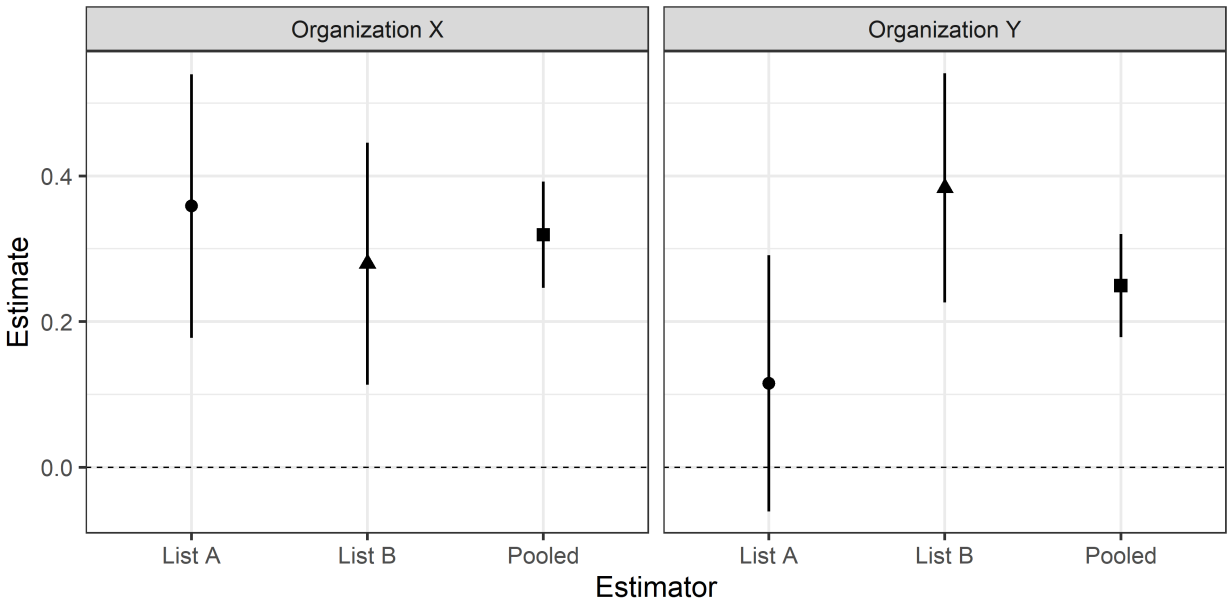

Figure 1: Standard and DLE estimates for Alvarez et al (2019)
Note: Rows indicate different estimators. Vertical lines denote 95 percent confidence intervals.

Y experiment overlap, one cannot determine whether this pattern reflects violations to the list experiment assumptions. The next section outlines two statistical tests that allow researchers to make uncertainty statements about this possibility.

3 Statistical tests

3.1 Setup

Previous work identifies two ways in which list experiment participants can alter responses in unintended ways (Zigerell 2011). If the sensitive item corresponds to an attitude that is

frowned upon, such as admitting to racial prejudice, then participants may deflate responses to avoid association with the sensitive item. Conversely, if the sensitive item is an attitude one would pretend to have, such as support for the regime in a dictatorship, then participants may react by inflating responses. Both behaviors can be true even for respondents who do not hold the sensitive item.[3]

The core intuition is that the opportunities to alter responses vary across treatment schedules. For simplicity, assume that respondents always see list A first and B second, this is true for the running example. Some DLEs shuffle the order of baseline lists. In this case, what matters for the statistical tests is whether the sensitive item appears first or second, regardless of baseline list order.

Let $Y_{iA} = z_i Y_{iA}(1) + (1 - z_i) Y_{iA}(0)$ be individual i's observed response to list A, and $Y_{iB} = (1 - z_i) Y_{iB}(1) + z_i Y_{iB}(0)$ the observed response to list B, with z_i indicating whether a respondent sees the sensitive item first. At the individual level, the researcher only observes the pair $(Y_{iA}(1), Y_{iB}(0))$ or $(Y_{iA}(0), Y_{iB}(1))$. These reflect treatment schedules with the sensitive item appearing first or second, respectively. Under the first schedule, respondents can react to the sensitive item in both questions. Under the second schedule, respondents can only react to the sensitive item in Y_{iB}.

To illustrate, imagine that participants deflate their response after seeing the sensitive item by one unit. Under the first treatment schedule, the relationship between $Y_{iA}(1)$ and $Y_{iB}(0)$ stays the same because both shift in the same direction. However, under the second treatment schedule, the paired responses become closer or further apart, depending on which one was larger before deflation. The goal of the tests is to detect this asymmetry in the aggregate.

[3] Another unintended response comes from non-strategic errors. I do not consider them here since previous work already outlines strategies to address them (Ahlquist 2017; Alvarez et al. 2019; Blair, Chou, and Imai 2019; Kuhn and Vivyan 2021; Riambau and Ostwald 2020).

3.2 Difference in differences

This test compares whether mean responses vary across treatment schedules. The quantities

$$\hat{\tau}_A = \frac{1}{N_{1A}} \sum_{i=1}^{N} z_i Y_{iA} - \frac{1}{N_{0A}} \sum_{i=1}^{N} (1 - z_i) Y_{iA} \tag{1}$$

and

$$\hat{\tau}_B = \frac{1}{N_{1B}} \sum_{i=1}^{N} (1 - z_i) Y_{iB} - \frac{1}{N_{0B}} \sum_{i=1}^{N} z_i Y_{iB} \tag{2}$$

denote the difference in means between responses with and without the sensitive item for the first and second list, with N_* as the sample size for the treatment and control groups in each question.

The null hypothesis is that the two differences in means are equal, $H_0 : \hat{\tau}_A - \hat{\tau}_B = 0$. For a DLE with fixed baseline list order, $\hat{\tau}_A$ and $\hat{\tau}_B$ correspond to the single-list prevalence estimates, and the test is equivalent to the consistency test proposed by Chuang et al. (2021). For the research design that shuffles baseline lists, this quantity is the difference in differences in means between responses to the first and second question instead.

Since the control group in the first question has not seen the sensitive item yet, the sign of the test statistic depends mainly on $\hat{\tau}_A$. A negative test statistic suggests deflation, while a positive value suggests inflation. Calculating the difference in differences is straightforward, but the computation of p-values must consider the clustered structure of the data, since each participant has two responses. Both randomization inference and OLS regression with clustered standard errors accommodate this structure.

3.3 Signed rank test

The alternative test evaluates whether one can attribute extreme differences in paired responses to the variation in treatment schedules. Rosenbaum (2007, 2020) proposes Stephenson's (1981) signed rank test to detect heterogeneous effects in pair randomized experiments.[4]

Since responses in a DLE are paired by participant, the test applies to the DLE setting. The test statistic is

$$\widetilde{T} = \sum_{i=1}^{N} \text{sgn}\{(z_i - (1 - z_i))(Y_{iA} - Y_{iB})\} \times \tilde{q}_i \tag{3}$$

which is the sum of Stephenson's signed ranks $\tilde{q}_i$, defined as

$$\tilde{q}_i = \binom{q_i - 1}{m - 1} \text{ for } q_i \geq m;\ \tilde{q}_i = 0 \text{ for } q_i < m \tag{4}$$

with q_i denoting the rank of the absolute difference in paired responses $|Y_{iA} - Y_{iB}|$. So $\tilde{q}_i$ records the number of possible subsets of size m in the data in which $|Y_{iA} - Y_{iB}|$ is the largest.

The choice of $1 \leq m \leq N$ determines what counts as an extreme difference. For example, with $m = 2$ the test is equivalent to Wilcoxon's signed rank, but with ranks ranging from 0 to $n-1$. As m increases, more ranks are considered zero and more weight goes to large differences. The choice of m is arbitrary, but researchers can calibrate its value at the pre-analysis stage.[5] In the application and simulations, I report $m = 10$ only to facilitate exposition. Section D of the appendix reports results under additional values.

Without ties in ranks, $\widetilde{T}$ is a distribution-free statistic, meaning its p-values are known in advance. With ties, one can compute exact p-values in small samples, while analytical derivation is a good approximation for experiments with large samples (Rosenbaum 2020).

[4] This is a general version of Wilcoxon's (1945) signed rank test.

[5] While $m = N$ is possible, it is not informative, since m should introduce meaningful variation in ranks.

To illustrate the behavior of this test statistic, consider a DLE with a zero prevalence sensitive item under no liars and no design effects. In this case, treatment assignment does not change responses, so the sign of $\tilde{q}_i$ flips randomly and $\widetilde{T}$ is zero in expectation. The only way $\widetilde{T}$ can be negative is in the presence of response deflation. However, $\widetilde{T}$ can be positive in the presence of at least one of response deflation or a non-zero prevalence rate. Section E of the appendix illustrates the underlying intuition.

This means that the signed rank test is more appropriate to evaluate the alternative hypothesis of $H_a : \widetilde{T} < 0$. Addressing response inflation with this test requires a null hypothesis different than the sharp null, which involves making statements on the prevalence rate, sample size, distribution of outcomes, and m. These are rarely known in advance.

3.4 Application

Table 2 applies both tests to the running example, treating each sensitive organization separately. Since Figure 1 suggests response deflation for Organization Y. I report two-sided p-values for the difference in differences and one-sided lower-tail p-values for the signed rank to avoid false positives.

For organization X, including the sensitive item in the first question leads to a difference in means about 0.08 points larger than in the second question, the p-value of 0.62 suggests little evidence against the null of equal differences in means. For Organization Y, the difference in differences is around -0.26, which implies a smaller difference in means when the sensitive item goes first. The p-value of 0.08 gives evidence against the null, although not sufficient to reject it under conventional standards.

The signed rank test statistic is positive for both sensitive items, and since both p-values are 1, one may conclude that the difference in differences test more appropriate. The simulations in the next section check if this intuition generalizes.

Table 2: Testing for response deflation in Alvarez et al (2019)

	Difference in differences		Stephenson's signed rank (m = 10)	
Experiment	Statistic	p-value	Statistic	p-value
Organization X	0.079	0.623	179.2×10^{21}	1
Organization Y	−0.268	0.082	182.6×10^{21}	1

4 Simulation

4.1 Setup

I simulate DLEs with a sample size of $1,000$ respondents and fixed list order. The potential outcome for responses to list A is $Y_i A(0) \sim B(4, 0.5)$. This implies four baseline items, each applying to respondent i with probability 0.5. This creates a distribution responses centered around middle values, which mimics an attempt to avoid floor and ceiling effects. The potential outcome for list B, $Y_i B(0)$, follows the same distribution and correlates with $Y_i A(0)$ with rank correlation ρ. I consider $\rho = \{0, 0.4, 0.8\}$ to capture how inducing correlations between lists affects performance.

I assume 15% of the respondents hold the sensitive trait at random. Following the simulations in Blair, Coppock, and Moor (2020), a single-list experiment is underpowered to detect this under conventional standards, but a DLE has over 80% power. This is a case in which opting for a DLE is consequential.

Also at random, a proportion $\gamma \in [0, 1]$ of the participants alter responses by 1 or 2 units when they see the sensitive item, doing so in both questions if they see it first. The magnitude is chosen at random and independently between questions. This reflects a setting where unintended responses are moderate but not symmetrical. To facilitate interpretation, I simulate response deflation and inflation separately. Figure D2 in the appendix shows how inflation and deflation introduce bias in estimates.

For each combination of parameters I simulate $1,000$ experiments and calculate power as the proportion of tests with p-values smaller or equal than 0.05. For the difference in means, the p-values are always two sided. For the signed rank, the p-values are left-tailed for deflation and right-tailed for inflation.

4.2 Results

Figure 2 shows the power of the different tests across parameter combinations for deflation and inflation. In general, power increases with the proportion of unintended responses in both tests. The exception is the signed rank test under inflation, which is sensitive to false positives as it captures the positive prevalence rate. Exception aside, both tests are well powered to detect a proportion of unintended responses that exceeds the true prevalence rate.

Everything else constant, the difference in differences has more power under response deflation than under inflation. One implication of this result is that, if possible, researchers should prefer sensitive items that are frowned upon over those one would pretend to have. For example, "I do not support the regime" is better than "I support the regime." Yet this conversion is not always straightforward.

Finally, as the correlation between baseline lists increases, the difference in differences has less power under both deflation and inflation. Moreover, under deflation, the performance of the signed rank test improves with the correlation. The difference appears trivial in stylized simulations, but Figure D4 in the appendix shows that it becomes more pronounced as the magnitude of response deflation increases. Since previous work recommends inducing positive correlation between lists to increase the precision of the DLE estimator (Glynn 2013), one should consider reporting both tests if response deflation is a concern.

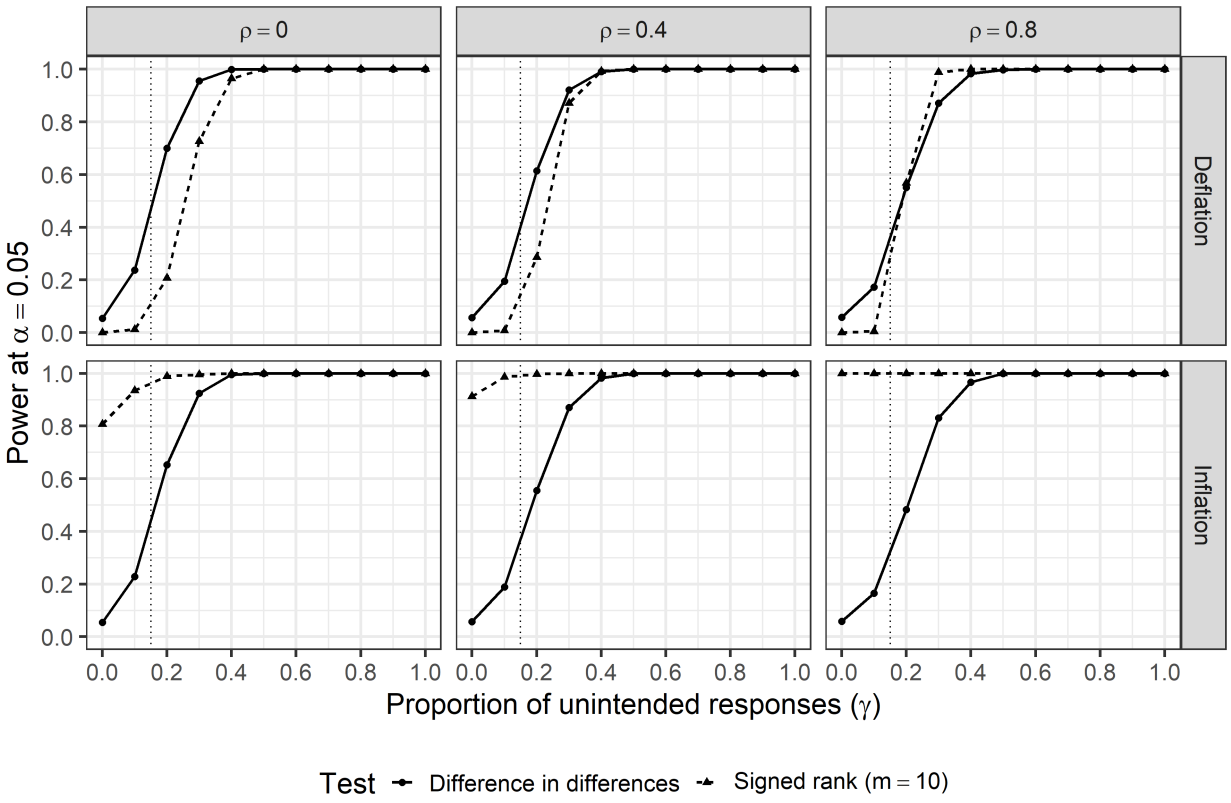

Figure 2: Statistical power under response deflation and inflation

Note: Each point is based on 1,000 simulations. The dotted vertical line denotes the true prevalence rate.

5 Conclusion

I propose two statistical tests to assess the validity of estimates in DLEs. This helps in facilitating the widespread implementation of a variant of the list experiment that improves along the bias-variance frontier. This is compatible with previous efforts to increase the precision , such as using responses to direct questions (Aronow et al. 2015) and auxiliary information (Chou, Imai, and Rosenfeld 2017) to adjust estimates accordingly.

These tests are most useful at the pre-analysis stage, as researchers can use them to calibrate baseline and sensitive items. Future work should use them as metrics to identify best research design practices to further improve our ability to address sensitive attitudes and behaviors through surveys.

References

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

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27 (2): 145–62.

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.

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.

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

Chuang, Erica, Pascaline Dupas, Elise Huillery, and Juliette Seban. 2021. "Sex, Lies, and Measurement: Consistency Tests for Indirect Response Survey Methods." *Journal of Development Economics* 148 (January): 102582.

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.

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.

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.

Kuhn, Patrick M., and Nick Vivyan. 2021. "The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." *Political Analysis*, April, 1–22.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the "New South"." *The Journal of Politics* 59 (2): 323–49.

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

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.

Rosenbaum, Paul R. 2007. "Confidence Intervals for Uncommon but Dramatic Responses to Treatment." *Biometrics* 63 (4): 1164–71.

———. 2020. *Design of Observational Studies*. Springer International Publishing.

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.

Stephenson, W. Robert. 1981. "A General Class of One-Sample Nonparametric Test Statistics Based on Subsamples." *Journal of the American Statistical Association* 76 (376): 960–66.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1 (6): 80.

Zigerell, L. J. 2011. "You Wouldn't Like Me When i'm Angry: List Experiment Misreporting." *Social Science Quarterly* 92 (2): 552–62.