Assessing the Validity of Prevalence Estimates in Double List Experiments*

Gustavo Diaz†

November 17, 2021

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 the estimator that follows from this indirect questioning technique is underpowered to capture the extent of sensitivity bias in common applications of interest. 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 as a measure of the prevalence of the sensitive trait in the population of interest. To overcome this difficulty, this paper proposes a statistical test to assess the validity of DLE estimates based on the difference in correlations between treatment schedules. I illustrate the implementation of this test with data from a previously published study on support toward anti-immigration organizations in California and further explore its properties via simulation.

*I thank Alex Coppock, Luke Sanford, and audiences at the 2020 NYU CESS Conference on Experimental Political Science, Polmeth 2021, and APSA 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 among a population of interest, with topics of interest 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 of the cumulative evidence in political science shows that the estimation technique that follows from the standard list experiment design is underpowered to capture the extent of sensitivity bias in common applications of interest. This happens because the bias reduction of list experiments comes at the cost of increased variance in prevalence estimates (Blair, Coppock, and Moor 2020).

The literature proposes double list experiments (DLEs) as an alternative research design to improve along the bias-variance frontier. Double list experiments 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, half of the variance of the estimate in the standard list experiment design (Droitcour et al. 1991).

While DLEs promise more precise estimates ny just adding an extra survey question, they are yet to become widespread practice. This is because this version of the research design brings the additional burden of choosing two comparable lists of baseline items (Glynn 2013). If the two baseline lists are not comparable, then the researcher cannot rule out the possibility of prevalence estimates emerging from unexpected respondent behavior, in which case the DLE estimator is not valid. This is a concern considering research that shows how using different baseline lists often yields diverging prevalence estimates of the same sensitive behavior (Chuang et al. 2021), and how prevalence estimates can vary in subsequent replications of the same list experiment (Gosen et al. 2018).

This paper proposes a statistical test to determine if the baseline lists in a DLE are comparable, which is tantamount to detecting joint violations in the list experiment identification assumptions of "no liars" and "no design effects" (Blair and Imai 2012). In practice, both assumptions imply that respondents do not strategically alter their responses when presented with the sensitive item. For example, respondents may deflate their responses to avoid association with a sensitive item that stands out (Zigerell 2011).

The test leverages variation in the timing with which the sensitive item is presented to respondents in double list experiments. I refer to these different timings as treatment schedules. In the usual DLE design, respondents see the two baseline lists close to each other in the survey flow and the sensitive item is 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 to both lists across treatment schedules, one can detect potential violations to the identification assumptions, which helps in assessing the validity of prevalence estimates.

More formally, I propose a statistical test based on the difference in correlations between treatment schedules. In principle, the researcher can use any masure of association of choice, yet I consider and evaluate three measures of correlation with different advantages and limitations: Pearson's correlation, Kendall's rank correlation, and distance correlation (Székely and Rizzo 2014).

I illustrate the application of the test with a reanalysis of a DLE on support for anti-immigration organizations in California (Alvarez et al. 2019) and examine its properties via simulation. The main takeaway from these exercises is that the test perform better when researchers induce positive correlation between baseline lists. This is yet another benefit of a practice that previous research shows to further reduce variance in prevalence estimates (Glynn 2013).

The key contribution of this paper is to facilitate the widespread implementation of DLEs by helping researchers to assess the validity of the prevalence estimates that follow from this research design. This, in turn, means an improvement along the bias-variance frontier in the ability to estimate sensitivity biases in the social sciences.

This paper also extends previous work aimed at detecting violations to the list experiment assumptions. Previous tests apply exclusively to the standard list experiment design (Blair and Imai 2012) or depend on the availability of direct responses to the sensitive item (Aronow et al. 2015). In the domain of DLEs, Chuang et al. (2021) propose consistency tests to assess comparability based on single-list point estimates. The test proposed here improves upon this by avoiding reliance on single-list estimates that usually exhibit high variability.

This paper is also compatible with previous efforts to increase the precision of prevalence estimates. For example, previous work uses responses to direct questions (Aronow et al. 2015) and axiliary information (Chou, Imai, and Rosenfeld 2017) to identify those who openly admit to the sensitive item and adjust estimates accordingly, Li (2019) introduces estimate bounds that allow relaxing the no liars assumption, and Riambau and Ostwald (2020) propose placebo-controlled list experiments as an alternative design that reduces inflation in prevalence estimates emerging from differences in list length.

2 List Experiment Designs

2.1 The standard list experiment

As a running example, consider the study conducted by Alvarez et al. (2019) on support for anti-immigration organizations in California.[1] Participants in an online survey in 2014 were asked the following question (emphasis and parentheses in original):

`Below is a list with the names of different groups and organizations on it.`

[1] This study also appears in Li (2019).

After reading the entire list, we'd like you to tell us how many of these groups and organizations you broadly support, meaning that you generally agree with the principles and goals of the group or organization. Please don't tell us which ones you generally agree with; ONLY TELL US HOW MANY groups or organizations you broadly support. HOW MANY, if any, of these organizations do you broadly support?

- Californians for Disability (organization advocating for people with disabilities)
- California National Organization for Women (organization advocating for women's equality and empowerment)
- American Family Association (organization advocating for pro-family values)
- American Red Cross (humanitarian organization)

In the standard list experiment, the control group sees the question as it appears above. The treatment group sees the following item included to the baseline list:

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

Respondents saw the name of a real organization, but the replication materials censor it for ethical reasons. The goal of the list experiment is to estimate the proportion of the target population who support this group. The difference in means of the responses in the treatment and control group is a valid estimate of the prevalence of the sensitive trait under standard experimental assumptions, plus two more (see Blair and Imai 2012; Glynn 2013 for details). The first assumption is that respondents do not misreport holding the sensitive trait. This is the "no liars" assumption. This assumption would be 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.

The second assumption is that participants do not alter their response to the baseline list when the sensitive item is included. This is the "no design effects" assumption, which is usually violated when respondents inflate or deflate their responses to avoid association with the sensitive item (Blair 2015; Zigerell 2011). Blair and Imai (2012) propose a test to detect potential violations of the no design effects assumption in the standard list experiment based on the comparison of marginal distribution of responses between treatment and control groups.[2]

While distinct in theory, these assumptions are hard to distinguish in practice. In both of the violations described above, the result would be a downward bias in the difference in means. In the extreme case, researchers can quickly identify violations if they end up with a negative prevalence estimate. In a trickier case, researchers risk a false negative in the form of not being able to detect non-zero prevalence when they should.[3]

A recent meta-analysis shows that the difference in means estimator in the standard list experiment is usually underpowered to detect sensitivity biases in common political science applications. This is because the bias reduction from list experiments relative to direct questions comes at the cost of increased estimate variance (see Blair, Coppock, and Moor 2020 for details).

2.2 Double list experiments

An underexplored solution to reduce the variability of list experiment estimates without compromising bias reduction is to implement a double list experiment (DLE) design (Droitcour et al. 1991). A DLE diverges from the standard list experiment in two ways. First,

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

[3]Non-strategic errors may result in violations that do not stem for response deflation. See Ahlquist (2017), Alvarez et al. (2019), and Blair, Chou, and Imai (2019) for strategies to address these.

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 with the following baseline items:

- `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, denote the list in the previous sub-section as list A and the current as list B. The second way in which DLEs differ from the standard design is that the sensitive item (Organization X) is randomly assigned to appear in list A or list B. This is equivalent to conducting two parallel list experiments. Some respondents receive list A under treatment and list B under control, others receive list A under control and list B under treatment.

This implies two difference in means estimates based on each list. The estimator in a DLE is the average of the two single-list difference in means. Moreover, while the variance of the standard list experiment estimator is the same as conventional difference in means, the variance of the DLE estimator is

$$V(\hat{\pi}_{DLE}) = \frac{1}{2n}[V(\boldsymbol{y_A(1)}) + V(\boldsymbol{y_B(0)}) - 2 \times Cov(\boldsymbol{y_A(1)}, \boldsymbol{y_B(0)}) + V(\boldsymbol{y_B(1)} + V(\boldsymbol{y_A(0)}) - 2 \times Cov(\boldsymbol{y_B(1)}, \boldsymbol{y_A(0)})] \quad (1)$$

Where V denotes variance and Cov the covariance, $\hat{\pi}_{DLE}$ is the DLE estimator, and n is the

sample size. $\boldsymbol{y_A(1)}$ and $\boldsymbol{y_A(0)}$ correspond to vectors of observed responses to list A under control and treatment, respectively, with the analogous for list B (See Droitcour et al. 1991; Glynn 2013 for further details).[4]

Because each respondent serves as both treatment and control in parallel experiments, DLEs produce estimates with roughly half of the variance than the standard DLE estimator (Blair, Coppock, and Moor 2020; Droitcour et al. 1991; Glynn 2013). This translates to more precise estimates when compared to the standard design. These estimates are valid as long as the list experiment assumptions hold.

DLEs promise reduction in the variance of prevalence estimates at the cost of just one additional survey question. However, the researcher must now justify the list experiment identification assumptions for two lists. Glynn (2013) suggests choosing baseline lists that correlate positively, which also has the benefit of further reducing variance in prevalence estimates. In practice, this entails constructing lists with similar items in terms of their applicability to respondents. For example, supporters of the California National Organization for Women (list A) are likely to also support Equality California (list B).

The challenge with choosing comparable baseline lists is that the same sensitive item can yield different single-list prevalence estimates. Recent work on the application of list experiments in the context of sexual behaviors in the Global South suggests this is a common result (Chuang et al. 2021). Different single-list estimates of the same sensitive item with comparable baseline lists hints at unexpected respondent behavior that violates the list experiment assumptions, which in turn raises concerns over whether the researcher can credibly aggregate the two single-list estimates into the pooled DLE estimate.

To make matters worse, researchers can find a scenario in which single-list point estimates are different, but the wide confidence intervals prevent them from determining whether estimates

[4] I use lowercase to denote vectors of outcomes observed in a subset of the population. For example, $\boldsymbol{y_A(1)}$ is the vector of responses to list A among those who were assigned to see the sensitive item in that list.

are different enough to worry. The design in Alvarez et al. (2019) provides an opportunity 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 studies. Since respondents always see list A first, the experiment has four possible combinations of sensitive items and their placement, these appear in Table 1. For each experiment, we can compute three different estimates: The two single-list experiment estimates for lists A and B, and the pooled DLE estimate.

Figure 1 shows the three 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 in 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.[5]

The study has comparable baseline items in list A and B, so the differences in estimates may come from the sensitive items. Organization Y seems more extreme than X, as they are group attempting to make matters against undocumented immigration on their own hands. The question is whether the different single-list estimates for Organization Y would emerge by chance or because the sensitive item stands out in a way that makes respondents deviate from the intended behavior.

[5]This does not imply that estimates for Organization Y are distinguishable from each other.

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

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

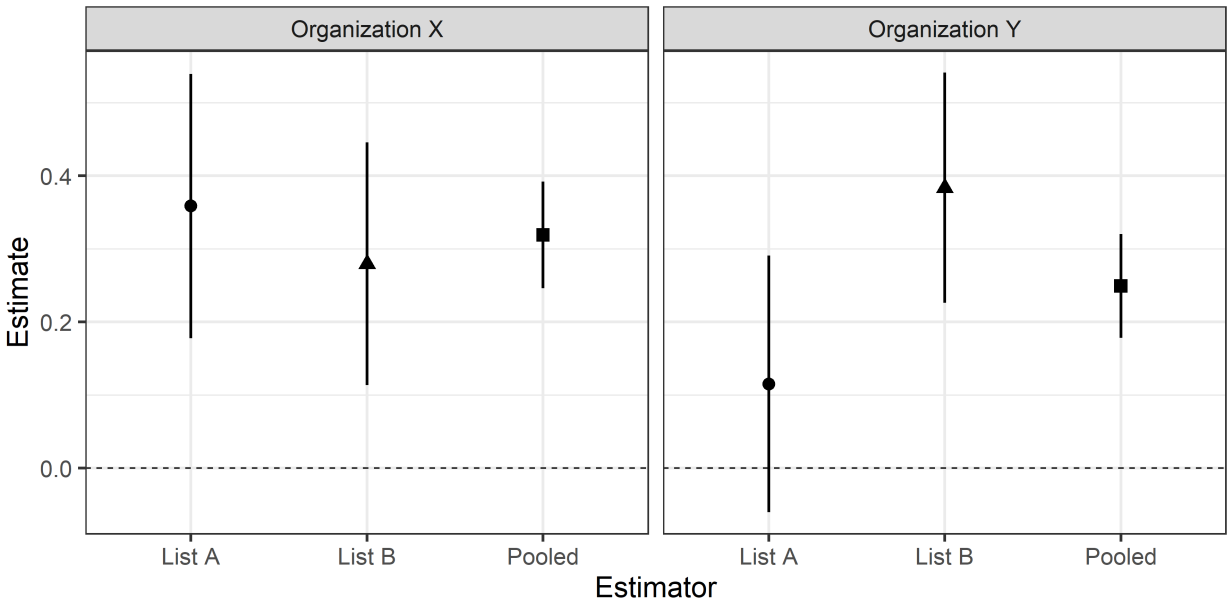

Figure 1: Standard and DLE estimates for Alvarez et al (2019)
Note: Each panel denotes effects for a different sensitive item. Rows indicate the corresponding standard and DLE estimators. Vertical lines denote 95 percent confidence intervals.

3 Assessing Estimate Validity

3.1 Difference in correlations test

I propose a statistical test that leverages variation in the treatment schedules of a DLE. A treatment schedule is the timing with which the sensitive item appears in the survey. In the conventional DLE design, the two baseline lists appear close to each other in the survey flow, and the researcher randomizes whether the sensitive item appears in the first or second list. What matters here is the order of appearance, even if the researcher shuffles which baseline list appears first. The test does not apply to designs that do not vary treatment

schedules. For example, the sensitive items in Chuang et al. (2021) always appear in the first list respondents see.

The core intuition of the test is to compare the relationship between responses to the two lists across treatment schedules. For simplicity, assume that respondents always see list A first and list B second, this is true for the running example. 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. In this case, z_i indicates whether a respondent sees the sensitive item in list A, which is the same as seeing the sensitive item in the first list.

At the individual level, the researcher only observes the pairs $(Y_{iA}(1), Y_{iB}(0))$ or $(Y_{iA}(0), Y_{iB}(1))$. At the aggregate, the researcher observes the paired response vectors $(\boldsymbol{y_A(1)}, \boldsymbol{y_B(0)})$ and $(\boldsymbol{y_A(0)}, \boldsymbol{y_B(1)})$. The first pair denotes the responses to lists A and B among respondents seeing the sensitive item first, the second pair denotes responses when the sensitive item appears second.

I propose the test statistic:

$$d = r_{\boldsymbol{y_A(1)}, \boldsymbol{y_B(0)}} - r_{\boldsymbol{y_A(0)}, \boldsymbol{y_B(1)}} \tag{2}$$

Where r is a measure of association between the paired response vectors indexed by the subscript. The next sub-section discusses calculating d using different correlation indices.

To illustrate how the test statistic works, consider a hypothetical example. Glynn (2013) suggests inducing positive correlation across lists to reduce variance in the DLE prevalence estimate. In the extreme case where both lists are essentially the same and the sensitive item has zero prevalence in the sample, then $d = 1 - 1 = 0$.

In a more realistic case, the expected correlation before introducing the sensitive item should be the same across treatment schedules. Under the standard list experiment assumptions, a sensitive item with positive prevalence will affect $r_{\boldsymbol{y_A(1)}, \boldsymbol{y_B(0)}}$ and $r_{\boldsymbol{y_A(0)}, \boldsymbol{y_B(1)}}$ in equal

measure. If the sensitive item induces responses that violate assumptions and respondents who see it in list A remember it by the time they are presented with list B. Then $r_{\boldsymbol{y_A(1)},\boldsymbol{y_B(0)}} \neq r_{\boldsymbol{y_A(0)},\boldsymbol{y_B(1)}}$.

This is because under the first treatment schedule respondents have the chance to strategically react to the sensitive item and alter their response in both lists, whereas under the second treatment schedule they can only react in the second list. In practice, such a difference can emergence by random chance, so the task is to generate a measure of information against the null hypothesis of $d = 0$, which helps researchers in determining whether they can credibly combine the two single-list estimates into the pooled DLE estimate.

The main attractive of this test is that it improves upon current alternatives in the DLE setting. For example, Chuang et al. (2021) propose consistency tests based on the difference in single-list prevalence estimates. This is a straightforward way to identify potential violations of the list experiment assumptions, but likely to be underpowered since single-list prevalence estimates tend to have high variance. For example, Figure 1 shows different single-list estimates for Organization Y, but one cannot determine if they are distinguishable from each other.

The main limitation of the test is that it requires the researcher to assume that the sensitive item has the same correlation with both baseline lists. Otherwise, we can have $d \neq 0$ even if both no liars and no design effects hold, resulting in a false positive.[6] In practice, researchers cannot determine whether a test statistic different from zero can be attributed to a potential violation of the identification assumptions or just a sensitive item with different correlation across lists.

However, current best practices in research design can help researchers to justify this assumption (Glynn 2013). For example, researchers are already advised to craft baseline lists so that, in expectation, they are positively correlated with each other. In practice, this means

[6] I than Alex Coppock for pointing this out. See section A of the appendix for more details.

creating lists that are very similar to each other, like lists A in B in the running example. Similar lists are more likely to have the same correlation with the sensitive item.

Another best practice is to shuffle the order of lists in the survey flow to cancel out order effects. Since the test depends on the timing of introduction of the sensitive item, regardless of which list comes first, shuffling baseline lists also guarantees that the elements on both sides of equation 2 are equivalent, so their expected correlation with the sensitive item will be the same unless the standard list experiment assumptions are violated.

3.2 Measuring correlation

The previous subsection defines d in terms of an arbitrary measure of association r because researchers can use any measure of their choice. Here I discuss three alternatives.

First, researchers can use Pearson's correlation index, which has the advantage of familiarity and an analytic derivation of the difference in correlations through Fisher's z-transformation (Fisher 1928).[7] The limitation is that this approximation may not be appropriate if the data does not follow a bivariate normal distribution, which is likely the case for ordinal responses to list experiment questions (Blair, Imai, and Lyall 2014; Hawkins 1989).

A second option is to use a non-parametric measure of correlation, such as Kendall's rank correlation. Blair, Imai, and Lyall (2014) use this as an alternative to Pearson's correlation in the context of comparing and combining responses to list and endorsement experiments. The advantage of Kendall's correlation is that it recognizes that responses are ordered but have no interval properties. As a non-parametric test statistic, Kendall's correlation makes no functional form assumptions and one can compute p-values through randomization inference. However, this test statistic is sensitive to ties, which can happen frequently given the relatively limited distribution of responses in list experiment questions.

The third option is to compute a more general measure of dependence, such as distance

[7] See Zou (2007) for a more recent discussion and a bootstrap confidence interval approach.

correlation (Székely and Rizzo 2014). This is the correlation of the scalar product of two double-centered distance matrices. Distance correlation has the advantage of being sensitive to non-linear forms of dependency. However, while the measure ranges from independence (0) to complete dependency (1), it places the burden of understanding the exact kind of dependency and whether it is problematic on the researcher. Moreover, in the context of a test statistic composed by the difference between two correlations, equal values may not necessarily imply equal forms of dependency.

3.3 Application

Table 2 implements the difference in correlations test using Pearson, Kendall, and distance correlation with data from Alvarez et al. (2019). As a benchmark, I also compute the consistency tests proposed by Chuang et al. (2021). The test statistic for this test is the difference between the two single-list prevalence estimates. The interpretation of this test is more straightforward, but prone to false negatives since single-list estimates tend to have high variability. I calculate different test statistics for each sensitive item since they entail independent experiments and compute p-values via randomization inference with 1,000 permutations.

In general, the tests show little evidence against the null hypothesis of equal correlations. This is plausible, since both baseline lists are similar and the researchers introduced several attention checks to exclude participants that may have inadvertently violated assumptions.

Still, the exercise produces some insights. First, for both sensitive items, the p-values of the difference in correlations tests are smaller than those of the difference in estimates. This suggests that the test proposed in this paper is more sensitive to potential violations in the identification assumptions.

Second, from the intuition in Figure 1, we should expect differences in correlations to be more pronounced for Organization Y, as it produced more varied standard list experiment

Table 2: Difference in correlations tests for the experiment in Alvarez et al (2019)

Test	Statistic	p-value
Organization X		
Difference in estimates	0.079	0.984
Pearson's correlation	0.024	0.578
Kendall's correlation	0.007	0.841
Distance correlation	0.013	0.735
Organization Y		
Difference in estimates	-0.268	0.907
Pearson's correlation	0.014	0.749
Kendall's correlation	0.028	0.463
Distance correlation	0.020	0.612

estimates. This is true for the test statistics using Kendall's and distance correlation as underlying measures of association, but not for Pearson's correlation. Similarly, the p-values for the tests using Kendall's and distance correlation suggest more evidence against the null hypotheses of the difference in correlations being equal to zero. While one cannot rule out the possibility of these differences arising due to random chance, this pattern highlights the limitations of applying Pearson's correlation to ordinal outcomes with relatively narrow distributions. If possible, researchers should resort to more than one measure of association.

4 Simulation

4.1 Setup

The application to a previously conducted DLE yields valuable yet inconclusive insights. In this section, I present the results of stylized simulations to illustrate the properties of the difference in correlations test more clearly.

Consider a DLE conducted on a sample of $N = 1,000$. For simplicity, assume that list A denotes whichever list appears first, even if the researcher shuffles the order. The individual potential outcome for responses to list A is $Y_{iA}(0) \sim B(4, 0.5)$. This implies four baseline

items, each applying to respondent i with probability 0.5. This creates a distribution of responses centered around middle values, which follows previous advice to avoid floor and ceiling effects (Blair and Imai 2012; Kuklinski, Cobb, and Gilens 1997; Kuklinski et al. 1997). The potential outcome for responses to list B, $Y_{iB}(0)$, follows the same distribution and correlates with $Y_{iA}(0)$ with rank-correlation ρ. I consider $\rho = \{0, 0.4, 0.8\}$ to capture how the correlations between the two lists affects the performance of the test.

Let $U \sim N(0, 1)$ be a latent vector. A respondent holds the sensitive trait Y_i^* with probability corresponding to its value in the probability distribution function of U. This means that roughly half of the sample holds the sensitive trait. This prevalence rate should be easily detected in a DLE with 1,000 participants (Blair, Coppock, and Moor 2020). I assume that Y_i^* is uncorrelated to $Y_{iA}(0)$ and $Y_{iB}(0)$ for the sake of clarity. Section B in the appendix shows that introducing positive correlation between the sensitive item and baseline lists leads to similar conclusions with less pronounced differences on the role of ρ.

The observed outcomes for each list are $Y_{iA} = Zi[\delta(Y_{iA}(0) + Y_i^*)] + (1 - Z_i)Y_{iA}(0)$ and $Y_{iB} = \delta(Y_{iB}(0) + (1 - Z_i)Y_i^*)$ with Z_i being a binary indicator of whether the sensitive item appears in list A and $\delta \in [0, 1]$ a deflation factor. The difference in observed outcomes captures the idea that respondents only deflate responses to list A when the sensitive item appears there, but always deflate them in list B regardless of treatment status. I focus on deflation since violations to the list experiment assumptions usually entail respondents shrinking their responses to avoid association with an undesirable item (Blair 2015; Zigerell 2011).[8]

I assign Z_i across respondents via complete randomization. Once realized, outcomes are truncated to the nearest integer to preserve ordinal outcomes. I simulate 1,000 experiments for each combination of parameter values, leading to a total of 33,000 simulations.

[8]There is a more nuanced case wherein respondents inflate or deflate their responses only when nearly none or all of the baseline items apply to them, respectively, since answering truthfully would imply admitting to the sensitive item. Future work should address this possibility, but I expect this to be a difficult problem to address since the usual research design advice is to avoid extreme responses.

4.2 Results

To illustrate the consequences of response deflation, Figure 2 shows how decreasing values of δ increase the bias of all possible list experiment estimators across simulations. Smaller values of δ imply that respondents shrink their responses more. The increase in bias is more pronounced in the estimator for list A, since whether or not response deflation occurs depends on the treatment schedule. Conversely, the change is less pronounced for list B since the current simulation setup implies that respondents always shrink responses. As expected, the performance of the DLE estimator lies in between the two single-list estimator that it averages over.

Figure 3 shows the performance of the test by comparing the relationship between δ against the value of alternative test statistics using a local polynomial fit. To make test statistics using different measures of correlation comparable, the figure reports scaled absolute values. If the test works as intended, one should expect smaller values of δ to map with higher values along the vertical axis. A steeper curve suggests that the test statistic is more sensitive to changes in δ.

Overall, Figure 3 suggests that the difference in correlations test is more sensitive when the two baseline lists are highly correlated. This is because the test statistic becomes less noisy as the two baseline lists approximate each other. In the hypothetical case in which the two baseline lists are identical, the only source of variation in the test statistic is the change induced by δ. None of the alternative test statistics seems to outperform the others, although distance correlation appears more sensitive to changes in δ when ρ is low.

Overall, the simulation results highlight yet another benefit of following previous research design advice. Constructing baseline lists not only reduces the variance in DLE prevalence estimates (Glynn 2013), it also makes it easier to detect potential violations to the identification assumptions induced by the double-list setting.

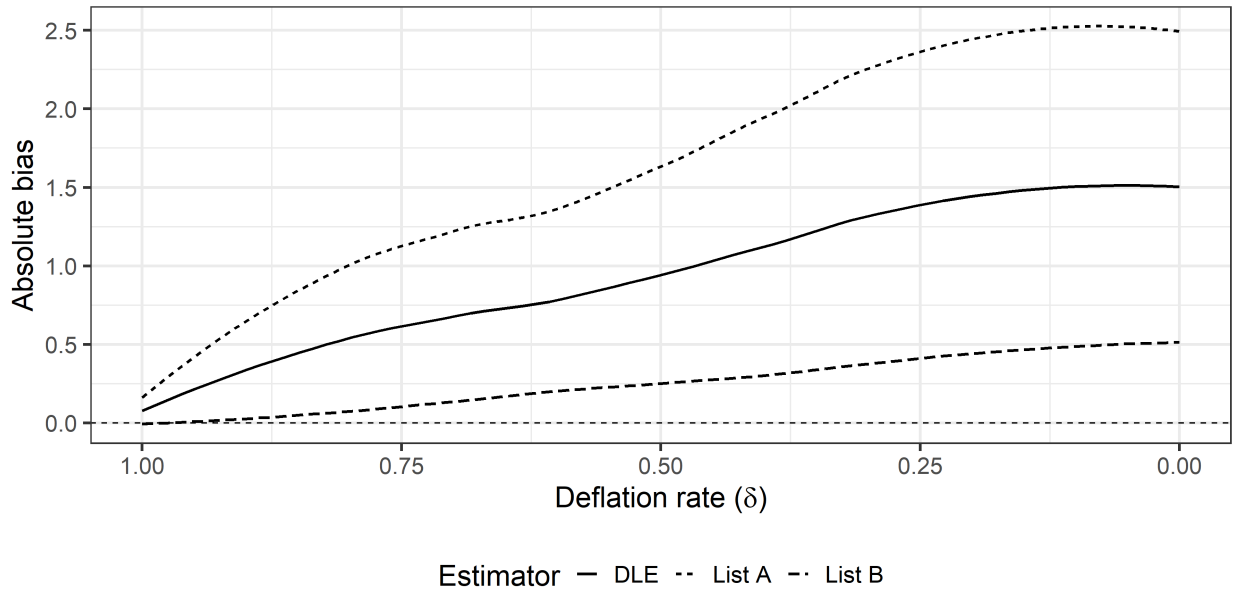

Figure 2: Deflation rate δ and estimator bias (LOESS fit)

5 Conclusion

I propose a statistical test to assess the validity of estimates in double list experiments (DLEs). This is a crucial step in facilitating the widespread implementation of the technique. DLEs improve along the bias-variance frontier by producing estimates with roughly half of the variance of single-list experiment estimates, but they require researchers to justify the identification assumptions of the technique for two baseline lists. Since applications in political science are generally underpowered to detect sensitivity biases, researcher should consider revisiting this underexplored variant of the list experiment.

I illustrate the usefulness of this test with a reanalysis of an experiment on support for anti-immigration organizations in California (Alvarez et al. 2019). This exercise shows that the test improves upon a direct comparison of single-list estimates, since these usually have wide confidence intervals. Using simulations, I also show that the test is more sensitive to response deflation, a common strategic behavior among respondents that violates identification assumptions, when the two baseline lists are highly correlated. This echoes previous research design advice on constructing comparable baseline lists, which already has the benefit of

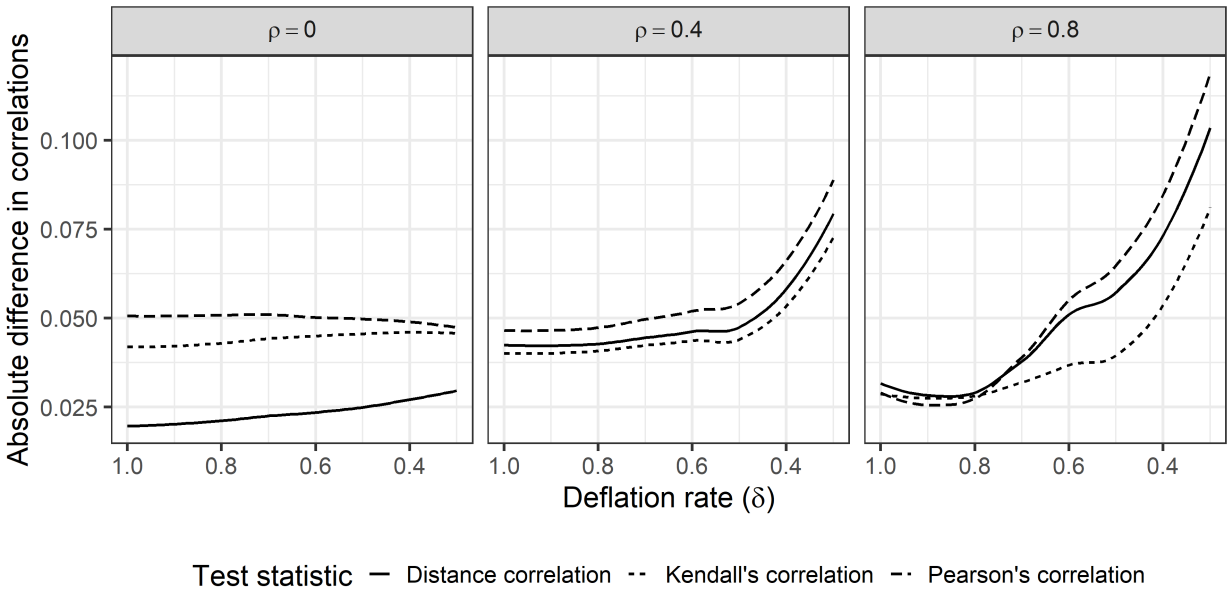

Figure 3: Difference in correlations test performance by test statistic and correlation between baseline lists ρ (LOESS fit)

further reducing the variance of DLE prevalence estimates (Glynn 2013).

The difference in correlations test is most impactful at the pilot stage, at which researchers need to compare the performance of alternative baseline lists, but may not have sufficient resources to detect anomalies by inspecting single-list estimates. Future work should use the test I propose as a benchmark to expand our knowledge on the consequences of key research design choices (e.g. shuffling list order, including placebo items) on respondents' strategic behavior.

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. https://doi.org/10.1017/pan.2018.57.

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

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. https://doi.org/10.1017/s0003055420000374.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77. https://doi.org/10.1093/pan/mpr048.

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.

Chou, Winston, Kosuke Imai, and Bryn Rosenfeld. 2017. "Sensitive Survey Questions with Auxiliary Information." *Sociological Methods & Research* 49 (2): 418–54. https://doi.org/10.1177/0049124117729711.

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. https://doi.org/10.1016/j.jdeveco.2020.102582.

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.

Fisher, Ronald A. 1928. "The General Sampling Distribution of the Multiple Correlation Coefficient." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 121 (788): 654–73. https://doi.org/10.1098/rspa.1928.0224.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum?" *Public Opinion Quarterly* 77 (S1): 159–72. https://doi.org/10.1093/poq/nfs070.

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. https://doi.org/10.1111/j.1540-5907.2011.00540.x.

Gosen, Stefanie, Peter Schmidt, Stefan Thörner, and Jürgen Leibold. 2018. "Is the List Experiment Doing Its Job?" In *Einstellungen Und Verhalten in Der Empirischen Sozialforschung*, 179–205. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-16348-8_8.

Hawkins, D. L. 1989. "Using u Statistics to Derive the Asymptotic Distribution of Fishers z Statistic." *The American Statistician* 43 (4): 235. https://doi.org/10.2307/2685369.

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. https://doi.org/10.1093/poq/nfp065.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the "New South"." *The Journal of Politics* 59 (2): 323–49. https://doi.org/10.1017/s0022381600053470.

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. https://doi.org/10.2307/2111770.

Li, Yimeng. 2019. "Relaxing the No Liars Assumption in List Experiment Analyses." *Political Analysis* 27 (4): 540–55. https://doi.org/10.1017/pan.2019.7.

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. https://doi.org/10.1017/psrm.2020.18.

Székely, Gábor J., and Maria L. Rizzo. 2014. "Partial Distance Correlation with Methods for Dissimilarities." *The Annals of Statistics* 42 (6). https://doi.org/10.1214/14-aos1255.

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

Zou, Guang Yong. 2007. "Toward Using Confidence Intervals to Compare Correlations." *Psychological Methods* 12 (4): 399–413. https://doi.org/10.1037/1082-989x.12.4.399.