Alexander Coppock*

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

DOI 10.1515/spp-2016-0005

Abstract: Explanations for the failure to predict Donald Trump's win in the 2016 Presidential election sometimes include the "Shy Trump Supporter" hypothesis, according to which some Trump supporters succumb to social desirability bias and hide their vote preference from pollsters. I evaluate this hypothesis by comparing direct question and list experimental estimates of Trump support in a nationally representative survey of 5290 American adults fielded from September 2 to September 13, 2016. Of these, 32.5% report supporting Trump's candidacy. A list experiment conducted on the same respondents yields an estimate 29.6%, suggesting that Trump's poll numbers were not artificially deflated by social desirability bias as the list experiment estimate is actually lower than direct question estimate. I further investigate differences across measurement modes for relevant demographic and political subgroups and find no evidence in support of the "Shy Trump Supporter" hypothesis.

Polling-based forecasts of the 2016 US Presidential election indicated that Hillary Clinton was likely to win. In Wisconsin, Clinton was projected to win 49.6% to 44.3%; instead she lost 46.5% to 47.2%, for a prediction error of 6 percentage points. In Michigan and Pennsylvania, the prediction errors were 4.5 and 4.4 points, respectively. While Clinton did win the popular vote 48.0% to 45.9%, she underperformed the pre-election prediction of 48.5% to 44.9%.[1] Explanations for this polling failure have included selection bias, faulty likely voter models, and measurement error. In this essay, I investigate the probable extent of a particular form of measurement error (the "Shy Trump Supporter" hypothesis) and argue that it is unlikely to be a major contributor to the evident error in pre-election forecasts of Clinton vote shares in many states.

[1] Estimates from fivethirtyeight.com's final pre-election forecast.

***Corresponding author: Alexander Coppock,** Yale University, New Haven, Connecticut 06520, USA, e-mail: alex.coppock@yale.edu. http://orcid.org/0000-0002-5733-2386

The "Shy Trump Supporter" hypothesis holds that polls understated support for Donald Trump because some respondents were reluctant to admit their support for his candidacy. This downward bias is hypothesized to be particularly pronounced on some phone surveys because they may have caused "shy" Trump supporters to dissemble to the live operators. The theoretical mechanism underpinning this possible source of error is social desirability bias. Donald Trump is often associated with socially-censured views on race and gender; some respondents who support him for partisan or policy reasons may not wish to appear racist or sexist to enumerators, others within earshot of the phone conversation, or even themselves. Social desiriability bias is hypothesized to impact responses not only in face-to-face or live telephone surveys but also in surveys administered online because subjects may wish to appear virtuous or otherwise correct even to unseen interviewers.

Social desirability may have biased polling estimates of support for candidates and policies prior to the 2016 election. The "Wilder Effect" (or Bradley Effect or Dinkins Effect) may occur if voters report support for Black candidates but do not actually vote for them (Hopkins 2009; Payne 2010). Polling misses in Britain have sometimes been attributed to "Shy Tories," voters who do not reveal their preference for the Conservative party to pollsters (Curtice 1997; Mellon and Prosser 2017). A similar logic has been applied to Canadian elections as well (Durand et al. 2001). Some have suggested that estimates of support for same-sex marriage were upwardly biased by social desirability (Powell 2013), though a list experiment approach similar to the one taken here (Lax et al. 2016) concludes that this is not the case.

Some evidence in favor of the Shy Trump Supporter hypothesis specifically comes from Dropp (2015), which documented an approximately six percentage point gap between estimates of Trump primary election support between live and online surveys. In a second study, Dropp found that the gap across survey mode narrowed substantially over the course of the general election campaign (Dropp 2016) though some survey mode differences remained among wealthier and college educated voters. A plausible explanation for this gap is social desirability bias, though differential nonresponse by survey mode (even after accounting for observable differences such as demographics) could also be the culprit.

An alternative strategy for measuring the extent of social desirability bias is a comparison of direct estimates of Trump support (obtained via standard vote preference questions) and indirect estimates obtained via list experiments (Miller 1984; Kuklinski et al. 1997; LaBrie and Earleywine 2000; Streb et al. 2008; Lyall et al. 2013; Frye et al. 2016). In a list experiment, subjects are first randomly assigned to treatment and control groups. Subjects in the control group are asked to report how many of a set of nonsensitive items they would do. Subjects in the treatment group are asked the same question, but the sensitive item is added to the list. Under the assumptions that subjects feel free to express themselves and that the number and composition of the list items do not affect responses, the list experiment yields estimates that are not biased due to social desirability. These list experiment assumptions are formalized in Blair and Imai (2012) as the No Liars and No Design Effects assumptions. If the list experimental estimate of Trump support exceeds the direct question estimate, we can conclude that some respondents indeed withheld their views from pollsters. If, on the other hand, direct questions and list experiments yield the same estimates of support, we can conclude that those who report not supporting Trump are sincere (provided the list experiment assumptions hold and not otherwise).

1 Design

Reuters/IPSOS conducted a nationally-representative online poll of 5290 adult Americans from September 2 to September 13. Trump support was first assessed with a direct question. The question read: "If the 2016 presidential election were being held today and the candidates were as below, for whom would you vote?" Respondents split 32.5% for Donald Trump and 37.0% for Hillary Clinton, while 30.5% reported that they would vote for other candidates, would not vote, or were still undecided. These estimates incorporate Reuters/IPSOS estimated sampling weights but do not condition on their likely voter screen.

The list experiment (administered well after the direct question was asked) sheds light on the question of whether 32.5% is an *underestimate* because some people are

Table 1: List Experiment Items.

Control List	Treatment List
If it were up for a vote, I would vote to raise the minimum wage to 15 dollars an hour	If it were up for a vote, I would vote to raise the minimum wage to 15 dollars an hour
If it were up for a vote, I would vote to repeal the Affordable Care Act, also known as Obamacare	If it were up for a vote, I would vote to repeal the Affordable Care Act, also known as Obamacare
If it were up for a vote, I would vote to ban assault weapons	If it were up for a vote, I would vote to ban assault weapons
	If the 2016 presidential election were being held today and the candidates were Hillary Clinton (Democrat) and Donald Trump (Republican), I would vote for Donald Trump.

ashamed to admit their support for Donald Trump. Subjects were randomly assigned to see the control or treatment version of the list experiment question, which read "Here is a list of [three/four] things that some people would do and some people would not. Please tell me HOW MANY of them you would do. We do not want to know which ones of these you would do, just how many. Here are the [three/four] things:"

The control and treatment list items are shown in Table 1. Following design advice in Glynn (2013), the control items were negatively correlated with one another to avoid floor and ceiling effects. Additionally, all list items were described as votes so that the presence of the putatively sensitive item (voting for Donald Trump) would not appear out of place.

2 Results

The list experimental results are reported in Table 2. The average number of items subjects reported doing in the control group was 1.548, whereas in the treatment group the average was 1.843. The difference in the averages forms the list experiment estimate of 0.296, or 29.6% support for Trump. Recall that the direct question estimate of Trump support was 32.5%, for a 2.9 percentage points difference. The bootstrapped standard error of this difference is 3.4 points, indicating that is not statistically significantly different from zero. If anything, the list experiment estimate is *lower* than the direct estimate, suggesting that (at least on average) the Shy Trump Supporter hypothesis is not supported in these data.

The list experiment and direct question estimates agree on average, but it is possible that social desirability may only affect a subset of voters. If any voters were shy about their support for Trump, they would likely be concentrated among those who are unwilling to admit their support outright. Among the 67.5% of the sample who did not state that they would vote for Trump when asked directly, the control group average was 1.610 and the treatment group average was 1.642.

Table 2: Distribution of List Experiment Responses by Treatment Condition.

	Control List	Treatment List
0 items	0.11	0.11
1 Item	0.37	0.22
2 Items	0.40	0.46
3 Items	0.13	0.15
4 Items		0.06
N	2645	2645

Entries are weighted proportions.

The list experiment therefore suggests that support for Donald Trump among this group was a mere 3.3%. This estimate is very small, and is itself not statistically significantly different from zero. The very low list experiment estimate among those who do not say they support Trump when asked directly is further evidence against the Shy Trump Supporter hypothesis.

Figure 1 extends the search for subgroups that hide their support for Donald Trump when asked directly to those formed by partisan affiliation, educational attainment, income quintile, gender, race or ethnicity, and vote propensity as modeled by IPSOS/Reuters. Four sets of estimates are presented. The unadjusted direct question and list experiment estimates do not condition on any covariates. The direct question estimate is the relevant sample mean and the list experiment

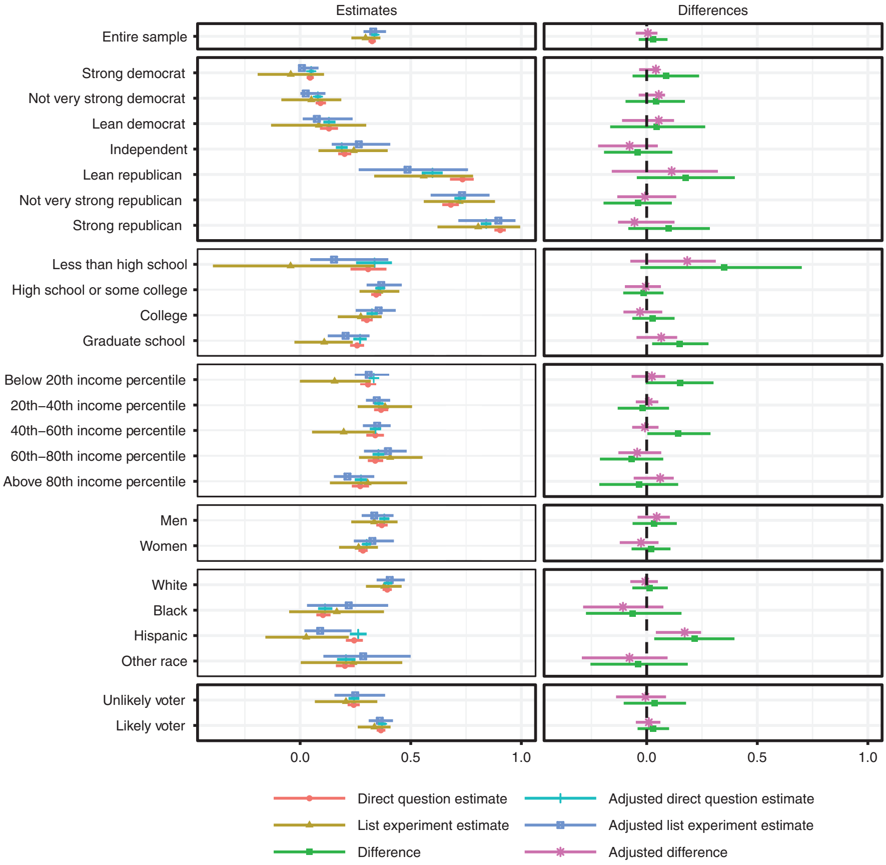

Figure 1: Direct and Indirect Estimates of Trump Support.

estimate is the relevant difference-in-means. I also present covariate-adjusted estimates. The response to the direct question is modeled using a logistic regression from which the average predicted probability of voting for Trump in the relevant subgroup is calculated. Using the nonlinear least squares (NLS) estimator proposed in Imai (2011), I generated covariate-adjusted list experiment predictions as well. In addition to making estimates more precise, the covariate-adjusted NLS estimator also has the virtue of constraining predicted probabilities to fall in the 0–1 range, something that the difference-in-means estimator does not guarantee. The response model for both the logistic and NLS regressions is given in Equation 1:

$$\begin{aligned} Y = {} & \beta_0 + \beta_1(\text{7-point Party ID}) + \beta_2(\text{Democrat}) + \beta_3(\text{Republican}) \\ & + \beta_4(\text{Less than High School}) + \beta_5(\text{High School}) + \beta_6(\text{College}) \\ & + \beta_7(\text{Black}) + \beta_8(\text{Hispanic}) + \beta_9(\text{White}) + \beta_{10}(\text{Female}) \\ & + \beta_{11}(\text{Likely Voter}) + \beta_{12}(22 - \text{point Income}) + \varepsilon \end{aligned} \tag{1}$$

A brief justification of this specification: all variables except for 7-point Party ID and 22-point Income are indicators, so not many modeling choices must be made. I allow some flexibility in the partisanship response curve by including indicators for being a Democrat or being a Republican, but not as much flexibility as would come with including indicators for all seven levels of partisanship. This model partially pools across subgroups; the estimated Trump support of Strong Republicans helps to estimate the Trump support of those who lean Republican. This model represents a compromise between no pooling across categories at all (the unadjusted models) and complete pooling (identical estimates for the entire sample). I will not present the coefficient estimates themselves, as they are not of any particular interest. Instead, I will present the average predictions that the model yields for the proportion of each subgroup that supports Donald Trump. The four estimators are not independent, so we must exercise caution when estimating the variances of the differences across estimators. For this reason, I will estimate standard errors and 95% confidence intervals for all quantities (including differences across estimators) via the nonparametric bootstrap.

The left-hand set of panels in Figure 1 presents all four estimates of Donald Trump support for each subgroup and the right-hand set of panels shows the differences between the direct question and list experiment separately for the adjusted and unadjusted estimators.[2] The estimates present a familiar pattern of

[2] For those who prefer tables, all estimates are reported in Appendix Table A1.

support: Democrats are unlikely to support Trump while Republicans are likely to do so. Overall, more educated groups support Trump less, though the estimates for those who have not graduated high school are uncertain as this group is relatively small. Different income groups did not diverge wildly in their support, nor did men and women. By any measure, Whites were more likely to support Trump than any other racial or ethnic category. Likely voters also expressed higher Trump support.

Differences across measurement mode, however, are few. Most of the 95% confidence intervals around the difference in the unadjusted estimates contain zero. We might be concerned that we fail to reject the null hypothesis that the two estimates are the same because the list experiment estimates are so uncertain. When we sharpen up the list experiment estimates with covariate-adjusted NLS regression, the confidence intervals do indeed shrink substantially, all but one (Hispanic) contain zero. While the possibility of social desirability affecting the direct responses of Hispanic voters in particular is theoretically interesting, a lone statistically significant difference out of 25 opportunities should not be overinterpreted. Broadly speaking, breaking down the sample by demographic subgroup does not uncover groups that appear to have systematically withheld their true views when asked the direct question.

3 Combined Estimate

Aronow et al. (2015) propose a method that combines direct questions and list experiments to form a more precise estimate of the prevalence of sensitive attitudes and behaviors. The estimator takes a weighted average of two quantities: the proportion of subjects who admit to supporting Donald Trump when asked directly and the list experiment estimate among those who do not admit supporting him when asked directly. In addition to the No Liars and No Design Effects assumptions required for standard list experiments, the combined estimator also requires an additional substantive assumption of “No False Confessions.” No False Confessions requires that subjects who do not support Donald Trump do not state that they in fact do support him when asked directly.

The combined estimate of Donald Trump support is 34.8% with a standard error of 2.9%. This estimate is 2.2 percentage points higher than the direct estimate, but this difference has a standard error of 2.9 percentage points, indicating that even with this more efficient estimator, we do not obtain a statistically significant difference between direct and indirect methods of measuring support for Donald Trump.

Aronow et al. 2015 also propose a joint test of the No Liars, No Design Effects, and the No False Confessions assumptions.[3] Under these three substantive assumptions, the list experiment estimate among those who admit to supporting Donald Trump should (in expectation) be equal to 1. Violations of any of these assumptions could lead the list experiment estimate to be different from 1. The list experiment estimate among those who directly admit supporting Trump is 85.7% with an estimated standard error of 5.6%. The two-sided *p*-value from a test of the null that the true parameter is 100% is 0.01, indicating that the joint test is failed in this case. One or more of the required assumptions is violated, so we should discount the value the combined estimate. As it happens, the difference between the combined and conventional list experiment estimates was small enough that our substantive conclusions about the plausibility of the Shy Trump Supporter hypothesis do not hinge on the credibility of the combined estimator.

4 Discussion

Observers trying to explain how the polls failed to predict a Trump win in 2016 have rightly considered how each assumption made along the road from raw polling data to vote share prediction could have been incorrect. The major classes of explanations concern how the set of people taking polls may differ from the population of interest (selection bias), how turnout decisions are modeled (faulty likely voter models), and whether standard horserace polling questions really measure what we hope they measure. The Shy Trump Supporter hypothesis falls into this last class of explanation: could it be that some people lie about their support because they are embarrassed?

The list experiment reported here suggests that this is not likely to be the case. No substantively meaningful differences across measurement mode were uncovered, even when considering special subgroups like independents or the college-educated. Subjects do not appear to be self-censoring when reporting that they do not support Trump.

The Shy Trump Supporter hypothesis first gained traction when there appeared to be a gap between online and phone survey estimates of Trump support during the primaries. It stands to reason that if online surveys are *free* from social desirability bias, then a list experiment conducted online (such as the study reported here) would be unlikely to uncover evidence of such bias. If so,

[3] The joint test also tests for violations of ignorability, but because the list experiment was administered after the direct question, this assumption is justified by design.

then we can downgrade the conclusions of this study to pertain to online surveys only. At a minimum, the list experiment evidence supports the conjecture that online polls do not suffer from this particular form of social desirability bias. However, online polls did not yield predictions of a Trump win any more than phone polls, so we are left to wonder which of the assumptions of the online polls were violated if not measurement fidelity.

As polling and academic communities consider recommendations for how to improve polling for future elections, good measurement should absolutely be top of mind. That said, the present study suggests that survey expressions of vote preference were not materially undermined by social desirability bias. If we consider measurement error to be a relatively minor source of prediction error, I think we should next turn our collective attention to the likely voter models that pollsters apply in order to better approximate the electorate. In my view, one step forward would be the adoption of data and analysis transparency. As a community, we need to be able to distinguish between the data that are generated by surveys and the models that translate them into predictions. Currently, the data and pollsters' beliefs (as reflected in their likely voter models) arrive as an undifferentiated bundle. As evidenced by the 2016 US Presidential election, aggregating these bundles into a polling average does not necessarily yield good estimates.

Appendix

Table A1: Comparing Direct Question and List Experimental Estimates of Trump Support.

Subgroup	N	Unadjusted estimates			Adjusted estimates		
		Direct question	List experiment	Difference	Direct question	List experiment	Difference
Entire sample	5290	32.5 (0.8)	29.6 (3.4)	3.0 (3.4)	33.6 (0.7)	33.0 (2.6)	0.6 (2.6)
Strong democrat	726	4.5 (0.8)	−4.3 (7.5)	8.8 (7.6)	5.0 (0.6)	0.8 (2.4)	4.2 (2.4)
Not very strong democrat	1140	9.2 (1.2)	5.0 (7.0)	4.2 (6.9)	8.0 (0.7)	2.5 (3.2)	5.5 (3.2)
Lean democrat	409	13.0 (2.1)	8.4 (11.0)	4.6 (11.1)	13.0 (1.4)	7.6 (5.9)	5.5 (6.1)
Independent	1131	20.1 (1.5)	24.3 (8.0)	−4.2 (8.1)	18.8 (1.4)	26.5 (6.9)	−7.7 (6.9)
Lean republican	351	73.4 (2.7)	55.9 (11.0)	17.6 (11.1)	59.8 (2.4)	48.5 (12.5)	11.3 (12.3)
Not very strong republican	990	68.1 (1.9)	72.1 (8.2)	−4.0 (8.0)	72.4 (1.3)	73.2 (6.9)	−0.8 (6.9)
Strong republican	543	90.4 (1.3)	80.5 (9.7)	9.9 (9.7)	84.1 (1.3)	89.6 (6.5)	−5.5 (6.6)
Less than high school	157	30.7 (4.2)	−4.3 (18.4)	35.1 (18.4)	33.6 (4.2)	15.3 (9.1)	18.3 (9.7)
High school or some college	2793	34.4 (1.1)	35.8 (4.7)	−1.4 (4.6)	36.2 (1.0)	36.7 (4.2)	−0.5 (4.2)
College	1465	30.1 (1.4)	27.4 (5.0)	2.7 (5.0)	32.5 (1.3)	35.5 (4.7)	−3.0 (4.6)
Graduate school	875	25.8 (1.7)	10.9 (6.7)	14.9 (6.5)	27.1 (1.6)	20.5 (4.7)	6.6 (4.5)
Below 20th income percentile	1038	30.7 (1.8)	15.6 (8.2)	15.1 (7.8)	33.3 (1.2)	31.0 (4.0)	2.4 (3.9)
20th–40th Income percentile	1361	36.6 (1.7)	38.4 (6.3)	−1.8 (6.2)	35.6 (1.1)	34.6 (2.8)	0.9 (2.7)
40th–60th Income percentile	936	34.0 (2.1)	19.7 (7.3)	14.3 (7.3)	34.0 (1.3)	34.8 (3.3)	−0.8 (3.1)
60th–80th Income percentile	1151	33.9 (1.8)	40.7 (7.4)	−6.8 (7.4)	35.4 (1.4)	39.7 (5.0)	−4.3 (4.9)
Above 80th income percentile	804	27.2 (2.0)	30.5 (8.8)	−3.4 (9.0)	27.5 (1.4)	21.4 (4.7)	6.1 (4.7)

Table A1 (continued)

Subgroup	N	Unadjusted estimates			Adjusted estimates		
		Direct question	List experiment	Difference	Direct question	List experiment	Difference
Men	2332	36.9 (1.3)	33.5 (5.3)	3.5 (5.2)	38.0 (1.1)	33.5 (3.7)	4.5 (3.7)
Women	2958	28.4 (1.1)	26.4 (4.4)	2.0 (4.5)	30.1 (1.0)	32.7 (4.6)	−2.6 (4.6)
White	3544	39.4 (1.1)	38.0 (4.1)	1.4 (4.1)	39.9 (0.9)	40.5 (3.3)	−0.6 (3.3)
Black	446	10.3 (1.7)	16.6 (10.9)	−6.3 (11.0)	11.2 (1.7)	22.0 (9.0)	−10.7 (9.0)
Hispanic	804	24.4 (2.0)	2.8 (9.7)	21.6 (9.4)	26.2 (1.9)	9.0 (5.5)	17.2 (5.4)
Other race	496	20.3 (2.2)	24.0 (11.6)	−3.8 (11.2)	20.7 (2.2)	28.5 (10.1)	−7.8 (10.0)
Unlikely voter	1420	24.2 (1.4)	20.7 (7.1)	3.5 (7.1)	24.3 (1.3)	24.9 (5.9)	−0.5 (5.9)
Likely voter	3870	36.5 (1.0)	33.6 (3.7)	2.9 (3.7)	37.0 (0.9)	36.0 (2.9)	1.0 (2.8)

All estimates incorporate sampling weights.
Bootstrapped standard errors are in parentheses.
Adjusted direction question estimates are predictions from a logistic regression.
Adjusted list experiment estimates are predictions from Imai’s (2011) NLS regression model.

References

Aronow, P. M., A. Coppock, F. W. Crawford and D. P. Green (2015) "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence," Journal of Survey Statistics and Methodology, 3:43–66.

Blair, G. and K. Imai (2012) "Statistical Analysis of List Experiments," Political Analysis, 20:47–77.

Curtice, J. (1997) "So How Well Did They Do? The Polls in the 1997 Election," Journal of the Market Research Society, 39:449–462.

Dropp, K. (2015) "Why Does Donald Trump Perform Better in Online Versus Live Telephone Polling?".

Dropp, K. (2016) "Yes, There Are Shy Trump Voters. No, They Won't Swing the Election," URL https://morningconsult.com/2016/11/03/yes-shy-trump-voters-no-wont-swing-election/.

Durand, C., A. Blais and S. Vachon (2001) "Review: A Late Campaign Swing or a Failure of the Polls? The Case of the 1998 Quebec Election," The Public Opinion Quarterly, 65:108–123.

Frye, T., S. Gehlbach, K. L. Marquardt and O. J. Reuter (2016) "Is Putin's Popularity Real?" Post-Soviet Affairs, 1–15.

Glynn, A. N. (2013) "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment," Public Opinion Quarterly, 77:159–172.

Hopkins, D. J. (2009) "No More Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead about Black and Female Candidates," The Journal of Politics, 71:769–781.

Imai, K. (2011) "Multivariate Regression Analysis for the Item Count Technique," Journal of the American Statistical Association, 106:407–416.

Kuklinski, J. H., P. M. Sniderman, K. Knight, T. Piazza, P. E. Tetlock, G. R. Lawrence and B. Mellers (1997) "Racial Prejudice and Attitudes Toward Affirmative Action," American Journal of Political Science, 41:402–419.

LaBrie, J. W. and M. Earleywine (2000) "Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique," Journal of Sex Research, 37:321–326.

Lax, J. R., J. H. Phillips and A. F. Stollwerk (2016) "Are Survey Respondents Lying About Their Support for Same-Sex Marriage? Lessons from a List Experiment," Public Opinion Quarterly, 80:510–533.

Lyall, J., G. Blair and K. Imai (2013) "Explaining Support for Combatants duringWartime: A Survey Experiment in Afghanistan," American Political Science Review, 107:679–705.

Mellon, J. and C. Prosser (2017) "Missing Nonvoters and Misweighted Samples: Explaining the 2015 Great British Polling Miss," Public Opinion Quarterly, forthcoming.

Miller, J. (1984) *A New Survey Technique for Studying Deviant Behavior*, Phd thesis., George Washington University.

Payne, J. G. (2010) "The Bradley Effect: Mediated Reality of Race and Politics in the 2008 US Presidential Election," American Behavioral Scientist, 54: 417–435.

Powell, R. J. (2013) "Social Desirability Bias in Polling on Same-sex Marriage Ballot Measures," American Politics Research, 41:1052–1070.

Streb, M. J., B. Burrell, B. Frederick and M. A. Genovese (2008) "Social Desirability Effects and Support for a Female American President," Public Opinion Quarterly, 72:76–89.