

Unresponsive and Unpersuaded: The Unintended Consequences of a Voter Persuasion Effort

Michael A. Bailey¹ · Daniel J. Hopkins² · Todd Rogers³

© Springer Science+Business Media New York 2016

Abstract To date, field experiments on campaign tactics have focused overwhelmingly on mobilization and voter turnout, with far more limited attention to persuasion and vote choice. In this paper, we analyze a field experiment with 56,000 Wisconsin voters designed to measure the persuasive effects of canvassing, phone calls, and mailings during the 2008 presidential election. Focusing on the canvassing treatment, we find that persuasive appeals had two unintended consequences. First, they reduced responsiveness to a follow-up survey among infrequent voters, a substantively meaningful behavioral response that has the potential to induce bias in estimates of persuasion effects as well. Second, the persuasive appeals possibly reduced candidate support and almost certainly did not increase it. This counterintuitive finding is reinforced by multiple statistical methods and suggests that contact by a political campaign may engender a backlash.

Keywords Field experiment · Political campaigns · Political persuasion · Non-random attrition · Survey response

✉ Daniel J. Hopkins
danhop@sas.upenn.edu

Michael A. Bailey
baileyma@georgetown.edu

Todd Rogers
Todd_Rogers@hks.harvard.edu

¹ Colonel William J. Walsh Professor of American Government, Department of Government and McCourt School of Public Policy, Georgetown University, Washington, DC, USA

² Department of Political Science, University of Pennsylvania, Philadelphia, PA, USA

³ Center for Public Leadership, John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA

Campaigns seek to mobilize and to persuade—to change who turns out to vote and how they vote. In many cases, campaigns have an especially strong incentive to persuade, since each persuaded voter adds a vote to the candidate's tally while taking a vote away from an opponent. Mobilization, by contrast, has no impact on any opponent's tally. Still, the renaissance of field experiments on campaign tactics has focused overwhelmingly on mobilization (e.g. Gerber et al. 2000, 2008; Nickerson 2008; Arceneaux et al. 2009; Nickerson and Rogers 2010; Sinclair et al. 2012; Rogers and Nickerson 2013), with only limited attention to persuasion.

To an important extent, this lack of research on individual-level persuasion is a result of the secret ballot: while public records indicate who voted, we cannot observe how they voted. To measure persuasion, some of the most ambitious studies have therefore coupled randomized field experiments with follow-up phone surveys to assess the effectiveness of political appeals or information (e.g. Adams and Smith 1980; Cardy 2005; Nickerson 2005a; Arceneaux 2007; Gerber et al. 2009, 2011; Broockman et al. 2014). In these experiments, citizens are randomly selected to receive a message—perhaps in person, on the phone, in the mail, or online—and then are surveyed alongside a control group whose members do not. Yet such designs have the potential for bias if the treatment influences participation in the follow-up survey.

This paper assesses one such persuasion experiment, a 2008 effort in which 56,000 Wisconsin voters were randomly assigned to persuasive canvassing, phone calls, and/or mailing on behalf of Barack Obama.¹ A follow-up telephone survey then sought to ask all subjects about their preferred candidate, successfully recording the preferences of 12,442 registered voters.

Focusing on the canvassing treatment, we find no evidence that the persuasive appeals had their intended effect. Instead, the appeals had two unintended effects. First, persuasive canvassing reduced survey response rates among people with a history of not voting. This result underscores a methodological challenge for persuasion experiments that rely on post-treatment surveys: persuasive treatments can induce differential attrition. To illustrate the potential for bias, we show that failure to account for treatment-induced selection in our data leads to demonstrably incorrect results when analyzing turnout.

Estimating treatment effects in the presence of attrition requires assumptions significantly stronger than those underpinning classical experimental analyses. In the spirit of Rubin and Schenker (1991), we thus estimate the persuasive effects of canvassing using various statistical approaches which vary in their underlying assumptions. Among those approaches is the method employed in most prior analyses of persuasion experiments, listwise deletion. Some of these approaches assume that responses to the follow-up survey are predictable from observed covariates while others do not. Regardless of the particular approach chosen, we uncover suggestive evidence of a second, unintended effect of canvassing: the pro-Obama canvass had a *negative* impact on Obama support of one to two percentage

¹ The data set and replication code are posted online at <https://dataverse.harvard.edu/dataverse/DJHopkins>. Due to their proprietary nature, two variables employed in our analyses are omitted from the data set: the Democratic performance in a precinct and each respondent's probability of voting for the Democratic candidate.

points. This backlash effect was statistically significant in many, but not all, specifications. As a consequence, we can rule out even small positive persuasive effects of canvassing with a reasonable degree of confidence.

This paper proceeds as follows. In section one, we discuss the literature on persuasion, focusing on studies that rely on randomized field experiments. We then detail the October 2008 experiment that provides the empirical basis of our analyses. In section three, we show how the experimental treatment affected whether or not individuals responded to the follow-up survey. To show how differential attrition can induce bias, we analyze voter turnout in the fourth section, contrasting the results based on the full sample with those for respondents to the phone survey. The non-random attrition produces a bias sizeable enough that a naive analysis of the survey respondents would lead one to mistakenly conclude that the canvass increased turnout. Turning to the analysis of persuasion, we present the estimated persuasion effects using models that embed different assumptions about the attrition.

In short, a brief visit from a pro-Obama volunteer made some voters less inclined to talk to a separate telephone pollster. It appears to have turned them away from Obama's candidacy as well. These results differ from other studies of political persuasion, both experimental (e.g. Arceneaux 2007; Rogers and Middleton 2015) and quasi-experimental (e.g. Huber and Arceneaux 2007). We conclude by summarizing the results and discussing ways in which they may or may not be generalizable.

Persuasion Experiments in Context

Political scientists have learned a great deal about campaigns via experiments (Green et al. 2008). The progress has been the most pronounced in the study of turnout, and for a straightforward reason: researchers can observe individual-level turnout from public sources, allowing them to directly assess efforts aimed at increasing turnout.

Still, there is more to campaigning than turnout. Campaigns and scholars care deeply about the effects of persuasive efforts. While there are various ways to study persuasion, a field experiment in which voters are randomly assigned to a treatment and then subsequently interviewed regarding their vote intention seems particularly attractive, offering the prospect of high internal validity coupled with a real-world political context.²

The motivation and design of such persuasion experiments draw heavily on turnout experiments, but differ in two important ways. First, it is quite possible that the campaign tactics which increase voter turnout may not influence vote choice. When people are mobilized to vote, they are being encouraged to do something that is almost universally applauded, giving inter-personal get-out-the-vote efforts the

² Strategies to study persuasion include natural experiments based on the uneven mapping of television markets to swing states (Simon and Stern 1955; Huber and Arceneaux 2007) or the timing of campaign events (Ladd and Lenz 2009). Other studies use precinct-level randomization (e.g. Arceneaux 2005; Panagopoulos and Green 2008; Rogers and Middleton 2015) or discontinuities in campaigns' targeting formulae (e.g. Gerber et al. 2011).

force of social norms (Nickerson 2008; Sinclair 2012; Sinclair et al. 2012). There is far less agreement on the question of *whom* one should support—and many Americans believe their vote choices to be a personal matter not subject to discussion (Gerber et al. 2013). It is quite plausible that voters may ignore or reject appeals to back a specific candidate, especially appeals that conflict with their prior views (Zaller 1992; Taber and Lodge 2006).

The conflicting findings of existing research on persuasion reinforce these intuitions. Gerber et al. (2011) find that television ads have demonstrable but short-lived persuasive effects. Arceneaux (2007) illustrates that phone calls and canvassing increase candidate support, and Gerber et al. (2011) and Rogers and Middleton (2015) show that mailings increase support. However, Nicholson (2012) concludes that campaign appeals do not influence in-partisans, but do induce a backlash among out-partisans, those whose partisanship is not aligned with the sponsoring candidate. Similarly, Arceneaux et al. (2009) show that targeted Republicans who were told that a Democratic candidate shared their abortion views nonetheless became less supportive of that Democrat. Nickerson (2005a) finds no evidence that persuasive phone calls influence candidate support in a Michigan gubernatorial race, and Broockman et al. (2014) find no evidence of persuasion through Facebook advertising. An experiment conducted with jurors in a Texas county concludes that attempts to apply social pressure can backfire (Matland et al. 2013), as can mis-targeted political appeals (Hersh and Schaffner 2013). In short, the evidence on persuasion effects is far more equivocal than that on face-to-face voter mobilization. Backlash effects are a genuine prospect (see also Bechtel et al. 2014).³

There is also the very real possibility that citizens' responsiveness to persuasion will vary with their political engagement. For instance, Enos et al. (2014) find that citizens who typically vote are more responsive to Get-Out-the-Vote efforts (see also Arceneaux et al. 2009). At the same time, Albertson and Busby (2015) show in a survey experiment that low-knowledge participants were demobilized by persuasive messages on climate change. Both findings are consistent with the idea that political engagement is a potentially important moderator. Political appeals might be off-putting to politically disengaged people, and backlash effects might be concentrated among that subset of the population. Moreover, as Enos et al. (2015) details, canvassers are likely to differ from the voters they are canvassing in consequential ways, a mismatch which may limit their effectiveness. Their differential interest in politics is but one such difference.

Persuasion experiments also differ from turnout experiments with respect to data collection. Turnout experiments use administrative records which provide reliable and comprehensive individual-level data. Persuasion studies, on the other hand, depend on follow-up surveys, with response rates of one-third or less being typical (see, e.g., Arceneaux 2007; Gerber et al. 2009, 2010, 2011). By the standards of contemporary survey research, such response rates are high. Still, there is little doubt that who responds is not random. In fact, Vavreck (2007) and Michelson

³ In a related vein, Shi (2015) finds that postcards exposing voters to a dissonant argument on same-sex marriage reduce subsequent voter turnout.

(2014) provide evidence that treatment effects differ when comparing the population of survey respondents to broader populations of interest. Given the high levels of non-response in prior studies of persuasion, differential sample attrition looms large as a possible source of bias.⁴

Wisconsin 2008

Here, we analyze a large-scale, randomized field experiment undertaken by a liberal organization in Wisconsin in the 2008 presidential election. Wisconsin in 2008 was a battleground state, with approximately equal levels of advertising for Senators Obama and McCain. Obama eventually won the state, with 56 % of the three million votes cast.

The experiment was implemented in three phases between October 9, 2008 and October 23, 2008. In the first phase, the organization selected target voters who were persuadable Obama voters according to its vote model, who lived in precincts that the organization could canvass, who were the only registered voter living at the address, and for whom the Democratically aligned data vendor Catalist had an address and phone number. By excluding households with multiple registered voters, the experiment aimed to limit the number of treated individuals outside the subject pool and improve survey response rates. Still, this decision has important consequences, as it removes larger households, including many with married couples, grown children, or live-in parents.

The targeting scheme produced a sample of 56,000 eligible voters. These voters are overwhelmingly non-Hispanic white, with an average estimated 2008 Obama support score of 48 on a 0 to 100 scale.⁵ The associated standard deviation was 19, meaning that there was substantial variation in these voters' likely partisanship, but with a clear concentration of so-called "middle partisans." Fifty-five percent voted in the 2006 mid-term election, while 83 % voted in the 2004 presidential election. Perhaps as a consequence of targeting single-voter households, this population appears relatively old, with a mean age of 55.⁶

In the second phase, every targeted household was randomly assigned to one of eight groups. One group received persuasive messages via in-person canvassing, phone calls, and mail. One group received no persuasive message at all, and the

⁴ Experimental studies also rely on self-reported vote choice, not the actual vote cast. This is less of a concern, as pre-election public opinion surveys like this one typically provide accurate measures of vote choice (Hopkins 2009).

⁵ Such support scores are commonly employed by campaigns. To generate them, data vendors fit a model to data where candidate support is observed, typically survey data. They then use the model, alongside known demographic and geographic characteristics, to estimate each voter's probability of supporting a given candidate in a much broader sample. The specific model employed is proprietary and unknown to the researchers. The Pearson's correlation with a separate measure of precinct-level prior Democratic support is 0.47, indicating the importance of precinct-level measures in its calculation in this data set. For more on the use of such data and scores within political science, see Ansolabehere et al. (2011), Ansolabehere et al. (2012), Rogers and Aida (2014) and Hersh (2015).

⁶ This age skew reduces one empirical concern, which is that voters under the age of 26 have truncated vote histories. Only 2.1% of targeted voters were under 26 in 2008, and thus under 18 in 2000.

other groups received different combinations of the treatments. The persuasive script for the canvassing and phone calls was the same; it is provided in the [Appendix](#). It involved an icebreaker asking about the respondent's most important issue, a question identifying whether the respondent was supporting Senator Obama or Senator McCain, and then a persuasive message administered only to those who were not strong supporters of either candidate.⁷ The persuasive message was ten sentences long and focused on the economy. After providing negative messages about Senator McCain's economic policies—e.g. “John McCain says that our economy is ‘fundamentally strong,’ he just doesn’t understand the problems our country faces”—it then provided a positive message about Senator Obama's policies. For example, it noted, “Obama will cut taxes for the middle class and help working families achieve a decent standard of living.” The persuasive mailing focused on similar themes, including the same quotation from Senator McCain.

Table 5 in the [Appendix](#) indicates the division of voters into the various experimental groups. By design, each treatment was orthogonal to all others. If no one was home during an attempted canvass, a leaflet was left at the targeted door. For phone calls, if no one answered, a message was left. For mail, an average of 3.87 pieces of mail was sent to each targeted household. The organization implementing the experiment reported overall contact rates of 20 % for the canvass and 14 % for the phone calls. It attributed these relatively low rates to the fact that the target population was households with only one registered voter. However, because leaflets and messages were left when voters were not reached, the actual fraction of voters who received at least some messaging was higher than the contact rates.

Our analyses operate within an intention-to-treat (ITT) framework. As is common in field experiments, not everyone answered the door when approached by a canvasser, just as not everyone watches television advertisements or reads campaign mail. It is quite plausible that people's availability is not random and, in fact, that some characteristics relevant to vote intentions may explain who among those randomly assigned to be visited actually talked to the canvassers (Gerber et al. 2012, p. 131). It is possible, for example, that those more enthusiastic about President Obama were more eager to talk to canvassers (whom they could likely guess were affiliated with a political campaign given that it was October in a presidential election year in a battleground state). If we were to compare the presidential vote intentions of those who actually talked to canvassers to the full control group, the groups would differ not only in that those who were treated were randomly selected for canvassing (which should be exogenous), but also in terms of factors that make them likely to talk to canvassers (which could be endogenous).

The ITT framework avoids this endogeneity by comparing the entire group of individuals assigned to treatment—whether they spoke to canvassers or not—to the entire group not assigned to be canvassed. The advantage of this approach is that we will not conflate treatment effects with factors associated with talking to canvassers. The downside of the ITT approach is that it will understate the true effect of

⁷ Specifically, voters were coded as “strong Obama,” “lean Obama,” “undecided,” “lean McCain,” and “strong McCain.”

canvassing on those who open their doors, as the treatment group will include some voters assigned to treatment but not actually treated. This makes the ITT a conservative estimand in some sense, and is a reason it is commonly used in instances in which some people do not “comply” with the treatment to which they were assigned (Gerber and Green 2012, Chapter 5). What is more, in a case like this in which we expect a field operation to canvass only some of the targeted voters, the ITT effect is itself a highly relevant quantity of interest. As it happens, the implementing organization did not provide individual-level information on who actually spoke to the canvassers, making the estimation of the ITT an obvious choice.⁸ Still, many experimental analyses of canvassing and other campaign tactics within political science report ITT estimates, sometimes alongside other causal estimands (e.g. Gerber et al. 2000; Nickerson 2005a; Huber and Arceneaux 2007; Arceneaux et al. 2009; Gerber et al. 2011; Broockman et al. 2014).

The randomization appears to have been successful. We assessed whether those assigned to the treatment and control groups differed on any of 18 covariates, including demographics such as age and imputed race alongside indicator variables for the number of prior elections in which each person voted. Even if there were no differences in treatment and control groups, it is possible that the treatment and control groups will differ simply by chance, given the large number of covariates we examine and concerns about multiple comparisons (e.g. Westfall and Young 1993). Therefore, we use an omnibus F test of the null hypothesis that no covariate predicts a variable indicating assignment to the canvass treatment. For the full sample, the F test has a p value of 0.19, indicating that jointly, the covariates are not strongly predictive of canvassing.⁹ That is exactly what we should expect given that canvassing was randomly assigned. Table 7 in the Appendix uses t tests for key covariates to further probe covariate balance in the full sample.¹⁰

In phase three, all targeted voters were telephoned for a post-treatment survey conducted between October 21 and October 23. In total, 12, 442 interviews were completed. To confirm that the surveyed individuals were the targeted subjects of the experiment, the survey asked some respondents for their year of birth, and 84 % of responses matched those provided by the voter file. The text of the survey’s introduction and relevant questions is provided in the Appendix.

Treatment Effects on Survey Response

If the treatment influenced who responded to the follow-up survey, any estimates from the subset of experimental subjects who responded are prone to bias. Accordingly, this section considers the impact of canvassing on survey response.

⁸ We can do additional analyses to approximate the effect of the treatment on people who actually spoke to the canvassers (the so-called Complier Average Causal Effect; see Angrist et al. 1996), and report the results in the Conclusion.

⁹ For the full regression, see the first column of Table 6 in the Appendix.

¹⁰ Similar results for the phone and mail treatments show no significant differences across groups.

For the full sample of 56,000 respondents, there are no pronounced differences between those who were canvassed and those who were not. But what about the smaller sample of 12,442 who responded to the survey? We again conduct an omnibus test by applying an F test to a regression of the canvassing treatment on 18 key covariates. Here, the corresponding p value is 0.006, indicating that whether people were canvassed is more strongly related to the covariates than expected by chance alone.¹¹

To probe the sources of that imbalance, Table 1 shows balance tests for subjects who completed the telephone survey. We highlight in bold those variables that have marked imbalances between voters assigned to canvassing and those not. Those who were assigned to canvassing were 2.0 percentage points more likely to have voted in the 2004 general election ($p = 0.001$), 3.4 percentage points more likely to have voted in the 2006 general election ($p < 0.001$), and 2.3 percentage points more likely to have voted in the 2008 primary ($p = 0.01$). It is important to note that the overall survey response rate was virtually identical for those assigned to canvassing and those not, at 22.2 %. Since these imbalances do not appear in the full data set of 56,000, these patterns suggest that canvassing changed the composition of the population responding to the survey.

Table 8 in the Appendix presents comparable results for the phone call and mailing treatments. There is some evidence of a similar selection bias when comparing those assigned to a phone call and those not. Among the surveyed population, 42.6 % of those assigned to be called but just 40.9 % of the control group voted in the 2008 primary ($p = 0.04$). For the 2004 primary, the comparable figures are 38.9 % and 37.3 % ($p = 0.07$). There is no such effect differentiating those in the mail treatment group from those who were not, suggesting that the biases are limited to treatments that involve interpersonal contact.

The relationship between being canvassed and subjects' decision to participate in the telephone survey appears related to their prior turnout history. In Fig. 1, we show the effect of canvassing on the probability of responding to the follow-up survey, broken down by the number of prior elections since 2000 in which each citizen had voted. Each dot indicates the effect of canvassing on the survey response rate among those with a given level of prior turnout. The size of the dot is proportional to the number of observations; the largest group is citizens who have voted in one prior election. The vertical lines span the 95 % confidence intervals for each effect.¹²

Among respondents who had never previously voted, the canvassed individuals were 3.9 percentage points *less* likely to respond to the survey. This difference is highly significant ($p < 0.001$). The effect is negative but insignificant for those who had voted in one or two prior elections. By contrast, for those who had voted in between three and six prior elections, the canvassing effect is positive, and for those who voted in exactly four prior elections, it is sizeable (2.9 percentage points) and statistically significant ($p = 0.007$). At the highest levels of prior turnout,

¹¹ For the corresponding regression model, see the second column of Table 6 in the Appendix.

¹² Voters under the age of 26 would not have been eligible to vote in some of the prior elections, and might be disproportionately represented among the low-turnout groups. We have age data only for 39,187 individuals in the sample. The negative effects of canvassing in the zero-turnout group persist (with a larger confidence interval) when the data set is restricted to citizens known to be older than 26.

Table 1 Balance among survey respondents

	Mean		<i>p</i> value	<i>N</i>
	Canvass assigned	Canvass not assigned		
Age	55.76	55.88	0.726	9,416
Black	0.017	0.018	0.671	12,442
Male	0.394	0.391	0.729	12,442
Hispanic	0.043	0.045	0.588	12,442
Voted 2002 general	0.242	0.232	0.163	12,442
Voted 2004 primary	0.390	0.371	0.031	12,442
Voted 2004 general	0.863	0.843	0.001	12,442
Voted 2006 primary	0.192	0.188	0.576	12,442
Voted 2006 general	0.634	0.600	0.000	12,442
Voted 2008 primary	0.429	0.406	0.011	12,442
Turnout score	3.263	3.149	0.005	12,442
Obama expected support score	47.36	47.95	0.100	12,440
Catholic	0.183	0.177	0.434	12,442
Protestant	0.467	0.455	0.181	12,442
District % Dem. 2004	54.66	54.86	0.353	12,440
District Dem. performance	58.01	58.18	0.374	12,440
District median income	46.26	45.94	0.155	12,439
District % single parent	8.19	8.28	0.212	12,439
District % poverty	6.22	6.40	0.127	12,439
District % college grads	19.79	19.58	0.279	12,439
District % homeowners	71.16	71.02	0.656	12,439
District % urban	96.64	96.96	0.099	12,439
District % white collar	36.31	36.29	0.882	12,439
District % unemployed	2.616	2.642	0.555	12,439
District % Hispanic	2.773	2.795	0.824	12,439
District % Asian	0.787	0.803	0.560	12,439
District % Black	1.849	1.878	0.759	12,439
District % 65 and older	22.82	22.80	0.921	12,439

This table uses *t* tests to report the balance between those assigned to canvassing and those not for individuals who completed the post-treatment phone survey

canvassing has little discernible influence on survey response, although these groups account for few individuals in the experiment.¹³

Here, too, it is reasonable to be concerned that these results might be a product of chance and appear only because of the number of sub-groups inherent in these analyses. Following Gerber et al. (2012, p. 301), we conducted an F-test in which we estimated a model of survey response on the full data set including interactions of the phone call and canvassing treatments with every turnout subgroup. We also

¹³ The effects for phone calls are generally similar, but not statistically significant (see Table 9 in the Appendix).

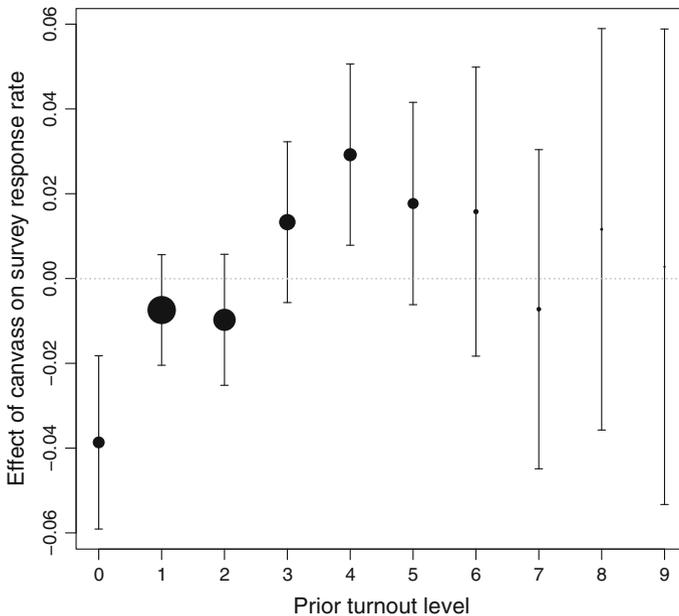


Fig. 1 Effect of canvass on survey response rates, by levels of prior turnout. Each dot indicates the mean effect, and its size is proportional to the number of citizens in that group. The vertical lines depict the 95 % confidence intervals.

included the 13 other covariates employed in the models below, meaning that the fully specified model had 31 total covariates. The F statistic from an analysis of variance comparing this model with a baseline model with only an intercept is 15.5, with a corresponding p value of less than 0.0001. In sum, the relationships detected here are far stronger than we would expect by chance, even accounting for the number of sub-groups being analyzed.

These results suggest that canvassing influences subsequent survey response in heterogeneous ways. It reduces the probability of survey response among those with low prior turnout while increasing the probability of survey response among those with middle levels of prior turnout. It is plausible that voters who infrequently vote and who are not strong partisans find such interpersonal appeals bothersome, and so avoid the subsequent telephone survey. With a canvasser on their doorstep, some individuals might feel pressure to remain in the conversation, even if they find the persuasive attempt off-putting. At the same time, the persuasive contacts in our experiment appear to trigger a pro-social response among those with middle levels of prior turnout. Such a response is consistent with prior research showing that those who vote inconsistently are the most positively influenced by mobilization efforts (Arceneaux et al. 2009; Enos et al. 2014), as ceiling effects limit mobilization's effects among the most likely voters.¹⁴

¹⁴ For example, Enos et al. (2014) find that direct mail, phone calls, and canvassing had small effects on turnout for voters with low probabilities of voting, large effects for voters with middle-to-high probabilities of voting, and smaller but still positive effects for those with the highest probabilities of voting.

Table 2 Breakdown of response differences

Sample	Mean Canvassed	Mean Control	Diff.	<i>t</i> test <i>p</i> value	<i>N</i>
Full Sample	0.318	0.318	0.000	0.861	56,000
Record of Outcome	0.336	0.335	0.001	0.634	41,808
+ Working Number	0.340	0.339	0.001	0.607	36,550
+ Participated in Survey	0.359	0.352	0.008	0.051	16,870
+ Reported Preference	0.363	0.350	0.013	0.005	12,442

This table reports the fraction of the previous nine elections in which respondents have voted, broken down by categories of response to the follow-up survey. The *p* values are estimated using two-sided *t* tests

To better understand the selection bias at work, it is important to identify precisely where in the survey process the systematic attrition appears. It turns out that the differences in prior turnout by canvass assignment are not due to differences in the ease of contacting voters. Table 2 shows the difference in the fraction of the prior nine primary and general elections in which the respondent voted between canvassed and non-canvassed subjects. The first row reiterates that when we compare all 28,000 respondents assigned to canvassing with the identically sized control group, there is essentially no difference in prior turnout between those assigned to treatment and control. There were 14,192 respondents whom the survey firm never attempted to call or who never answered the phone, providing no record of the outcome. But as the second row illustrates, the removal of those respondents leaves treatment and control groups that are well balanced in terms of prior turnout. Another 5,258 subjects had phone numbers that were disconnected or otherwise unanswerable—but the third row shows that there was little bias in prior turnout for the 36,550 cases where the phone rang and where we have a record of the subsequent outcome. The process of selecting households to call and calling them does not appear to have induced the biases identified above.

The fourth row in Table 2 shows that the sample drops by nearly half when restricted to the 16,870 respondents who were willing to participate in the survey. And here, there is evidence of bias, with the remaining members of the treated group having a prior turnout score 0.008 higher than the control group’s ($p = 0.051$). The bias grows further when examining the 12,442 respondents who actually reported a candidate preference, with the difference in prior turnout becoming 0.013 ($p = 0.005$). Being canvassed leads some higher-turnout respondents to be more likely to participate in the survey relative to the control group. Persuasion attempts have a demonstrable effect on who participates in an ostensibly unconnected survey in the following weeks.

Selection Bias and Turnout

Previous research has documented that experimental effects can differ for survey respondents (Vavreck 2007; Michelson 2014), raising concerns about external validity. Here, we have shown that experimental treatments can influence survey

Table 3 OLS estimates of effect of treatments on probability of turnout

	All subjects	Survey sample only
Canvass	0.003 (0.004)	0.015 (0.008)
Phone call	-0.004 (0.004)	0.013 (0.008)
Mail	0.001 (0.004)	-0.005 (0.008)
Constant	0.664* (0.004)	0.726* (0.008)
<i>N</i>	56,000	12,442

Standard errors in parentheses

* Indicates significance at $p < 0.05$

response, raising the possibility of bias and threatening even internal validity. But does that differential responsiveness to the survey affect our causal estimates? One way to assess this question is to look at turnout. From administrative data, we know the right answer, as we have data on turnout for all 56, 000 subjects. The column on the left of Table 3 uses a linear probability model to show that the canvass, phone, and mail treatments had no significant effect on turnout in the 2008 November election for the full sample.¹⁵ The difference between this null effect and the positive effects reported in much of the prior research (Green et al. 2008; Arceneaux et al. 2009) might be an indication that persuasive appeals can be off-putting.

If we look only at those who responded to the survey, however, we get a different answer. The right column of Table 3 shows the result from the same model estimated only for individuals who responded to the survey. Canvassing now appears to be associated with a 1.5 percentage-point increase in turnout ($p = 0.05$), and phone calls with an additional 1.3 percentage-point boost ($p = 0.09$). These ostensible effects are spurious and due entirely to selection. We know from the above discussion that the canvassing treatment had different effects on different groups: canvassing turned off people unlikely to vote from answering the follow-up survey while it encouraged people who vote sporadically. This means that in the survey sample, we have removed a disproportionate number of low-turnout voters who were canvassed and included a disproportionate number of moderate-turnout voters who were canvassed, thereby inducing a spurious association between canvassing and turnout.

The point is to demonstrate empirically that non-random attrition can matter. The experiment's sponsors did not intend the treatments to affect turnout, as they were designed to be persuasive. They were also administered to a sample selected to avoid strong partisans. Yet if we had been limited to only the surveyed sample and had analyzed that sample without considering the selection process, we would have inferred incorrectly that canvassing (and perhaps phone calls) increased turnout.

There are two important implications of the findings on survey responsiveness and voter turnout. First, the treatments did in fact induce behavioral responses—just

¹⁵ Results using logistic regression are highly similar.

not behavioral responses expected. Those individuals who were least inclined to vote responded to a persuasive canvassing visit by becoming markedly less likely to complete a seemingly unconnected phone survey. Second, this pattern of heterogeneous non-responsiveness raises the prospect of bias when assessing the primary motivation of the experiment: whether or not the persuasion worked. In the next section, we discuss estimating treatment effects in the presence of sample selection.

Estimating Treatment Effects on Vote Intention

The goal of the persuasion campaign was, of course, to increase support for Barack Obama. The statistical challenge is accounting for selection effects. Not only do we harbor the general concern that the sample of those who answered the follow-up survey is non-random, the previous section provided evidence that the canvassing treatment itself induced some low-turnout respondents to not respond to the survey while having the opposite effect among higher-turnout voters.

In the [Appendix](#), we provide a formal discussion of the conditions of selection bias. As that discussion makes clear, to recover treatment effects in the face of attrition which is related to treatment assignment, we need to invoke additional assumptions (Gerber et al. 2012; Little et al. 2012). Rather than choose and argue for a single set of assumptions, we employ two techniques that use differing assumptions to address the attrition present in this experiment—and detail the results of several other techniques in the [Appendix](#). Specifically, the first approach is a variant of multiple imputation, and it assumes data are missing at random conditional on the covariates observed. The second is a non-parametric selection model, which allows for non-ignorable missingness by instead making specific assumptions about the selection process, the outcome process, and the relationship between them.¹⁶

Multiple Imputation using Chained Equations. One technique for addressing missing data in both covariates and outcomes is multiple imputation (Schafer 1997; King et al. 2001; Little et al. 2002), a technique which uses observed covariates to provide information about a respondent's likely response had she completed the survey.¹⁷ Standard approaches to multiple imputation assume that the data are “missing at random,” meaning that conditional on the observed covariates, the pattern generating missing observations is random. Put differently, we are assuming that the missing data can be predicted with the observed covariates, including characteristics of the subjects themselves (e.g. age, prior vote history, gender, etc.) and their neighborhoods (e.g. percent Democratic, median household income, percent with a Bachelor's degree, etc.). How tenable that assumption is hinges on

¹⁶ In separate, ongoing research, we use the turnout results described above as a benchmark with which to evaluate each of these methods.

¹⁷ As Little et al. (2012) explain, “weighted estimating equations and multiple-imputation models have an advantage in that they can be used to incorporate auxiliary information about the missing data into the final analysis, and they give standard errors and p values that incorporate missing-data uncertainty”(1359).

the quality of the observed covariates. Still, unlike some methods, variants of multiple imputation can handle missingness across multiple variables with no added complexity, making them appropriate for a range of missing-data problems. It also includes listwise deletion—the approach employed in prior analyses of persuasion experiments—as a special case.

The approach to multiple imputation we employ is “Multiple Imputation using Chained Equations” (MICE) (Buuren et al. 2006). In contrast to other approaches, MICE involves iteratively estimating one variable at a time through a series of equations with potentially differing distributional forms. This procedure affords it greater flexibility in its handling of variables that are not continuous, such as the binary outcome of interest here.¹⁸

To address the varying survey responsiveness across prior turnout levels, our imputation and outcome models include a single, continuous measure of the number of prior elections in which the subject voted and 18 indicator variables interacting the canvassing and phone call treatments with each of the nine possible levels of prior turnout. We also include several other variables that could affect both whether and how an individual responded to the survey. We use Catalyst’s Democratic support score, a continuous measure which draws on various demographic and proprietary data sources. We control for gender, age, race, ethnicity, and religion. The race and religion variables are from Catalyst models predicting the likelihood a person is Black, Hispanic, Protestant or Catholic. We also use tract-level measures of the median income in the respondent’s neighborhood and the percentage of college graduates as well as a separate, composite measure of Democratic voting in the respondent’s precinct.

We impute outcome measures as well, a choice which induces no bias under the “missing at random” assumption. The outcome of primary interest is a binary indicator which is 1 for surveyed respondents who support Obama and 0 for those who are undecided or support McCain. 58 % of those who responded supported Obama, while 26 % supported McCain and 16 % were unsure. As is standard with multiple imputation, we impute possible values of each missing observation and then combine the analyses of these data sets.¹⁹

As a baseline, we first estimate a model using the 12,442 fully observed cases (which we refer to as the listwise deletion model given that any observation with any missing variable is deleted from the analysis). The estimated difference in Obama support between those who were canvassed and those who were not was –1.6 percentage points ($p = 0.13$, two-sided) controlling for the covariates listed above. This result suggests that if anything, canvassing made respondents *less* likely

¹⁸ But that fact also means that the “implied joint distributions may not exist theoretically” (Buuren et al. 2006, p. 1051). Still, that important theoretical limitation does not prevent MICE from working well in practice (Buuren et al. 2006).

¹⁹ To examine the performance of our model for multiple imputation, we performed tests in which we deliberately deleted 500 known survey responses from the fully observed data set ($n = 12,442$) and then assessed the performance of our imputation model for those 500 cases where we know the correct answer. In each case, we used the full multiple imputation model to generate five imputed data sets for each new data set, and then calculated the share of deleted responses which we correctly imputed. The median out-of-sample accuracy across the resulting data sets was 74.9 %, with a minimum of 73.3 % and a maximum of 76.0 %. This performance is certainly better than chance alone.

to report supporting Obama. Given the results on survey response above, it is possible that the effect is even more negative if Obama opponents were especially put off by the canvassing and, therefore, especially unlikely to respond to the survey.

The results of the imputation reinforce that possibility. We first estimate the treatment effect for all the imputed respondents by using logistic regression and then combining the estimates from the five data sets appropriately. For the full data set, the estimated treatment effect after multiple imputation is -1.88 , with a 95 % confidence interval from -2.94 to -0.79 percentage points. Under this model, the persuasion effect of canvassing for the overall population was *negative*, and significantly so.²⁰ When we remove the 11,125 subjects who had no phone match score and were thus harder to reach by phone, we find that the treatment effect declines very slightly to -1.71 .²¹

Given that canvassing had a negative effect on survey response among infrequent voters, it is valuable to examine canvassing's impact on Obama support among that same group. To do so, we fit a logistic regression similar to that described above to the imputed data sets with the 29,533 respondents who had turned out in no more than 2 of the prior 9 elections. Among that group, the estimated treatment effect nearly doubles, to -3.7 percentage points, with a 95 % confidence interval from -5.2 to -1.3 percentage points. Here, we see stronger evidence that canvassing is off-putting to infrequent voters: not only does it encourage them to avoid a subsequent survey, but it also makes them markedly less likely to support the candidate on whose behalf the persuasion was undertaken.

Non-Parametric Selection Model. Next we present results from the non-parametric, two-stage estimator for selection models detailed by Das et al. (2003). The key difference from multiple imputation approaches is that this estimator allows errors to be correlated across the selection and outcome equations. This particular estimator has a motivation similar to the two-stage Heckman estimator (Heckman 1976), although it is less reliant on a particular functional form assumption.

In the first stage, we model the probability of survey response for each respondent. In this model, we control for the same set of covariates as those described above for multiple imputation. We also use three additional variables which are related to the vendor-assessed quality of the phone number information: indicator variables for weak phone matches, medium phone matches, and strong phone matches (with no phone match being the excluded category). There is an exclusion restriction at work here. We are assuming that these factors predict whether or not someone answered the phone survey but do not, conditional on the other variables in the model, predict vote intention.

In the second stage, we then condition on various functions of the estimated survey response probability. Table 10 in the Appendix displays the second-stage results for multiple specifications of the non-parametric selection model. For the full

²⁰ In fact, the associated p value is less than 0.002, meaning that the finding would remain significant even after a Bonferroni correction for multiple comparisons to account for the analyses of the phone and mail treatments.

²¹ The associated 95 % confidence interval spans from -3.03 to -0.60 .

Table 4 Overview of all results

Missing data strategy	Lower bound	50th	Upper bound
Listwise deletion—no covariates	-3.64	-1.61	0.46
MICE, all observations	-2.94	-1.88	-0.79
MICE, phone score	-3.03	-1.71	-0.60
Non-parametric Selection	-3.95	-1.88	0.17
ABB, phone score, prior = 0, $k = 3$	-3.29	-1.65	-0.01
ABB, phone score, prior = -3.5, $k = 3$	-3.34	-1.73	-0.05
Inverse propensity weighting	-3.52	-1.79	-0.05
Heckman selection	-3.29	-1.55	0.01

This table reports the 2.5th, 50th, and 97.5th percentiles of the average treatment effect. The units are percentage points

Note “Phone score” refers to the 44,875 experimental subjects for whom a pre-treatment phone match score was available via Catalist. For the Approximate Bayesian Bootstrap (ABB), the prior indicates the level by which Obama support was adjusted in among unobserved respondents. As k increases, the preference for matching similar observations in the ABB increases.

sample, the effect of canvassing is negative at -1.9 percentage points, with a p value of 0.08. The second column of Table 10 shows that the effect is large, but slightly less certain ($p = 0.11$), when we examine only respondents who have turned out in fewer than 3 recent elections.

Alternate Estimators As detailed above, dealing with sample selection requires assumptions well beyond those justified by randomization alone. To demonstrate the consistency of the core results in the face of different assumptions, Table 4 summarizes the results across various methods for dealing with missing data employed both here and in the [Appendix](#).

The first four rows of the Table 4 present the results we have already discussed using listwise deletion, multiple imputation, and a non-parametric selection model. The additional rows summarize results we detail in the [Appendix](#). The fifth and sixth rows present results from Approximate Bayesian Bootstraps (ABB) (Siddique and Belin 2008b). A variant of hot-deck imputation, this approach can allow for non-ignorable missingness through the use of a prior on the outcomes of unobserved respondents. The fifth row (in which the prior is 0) reports the results when we assume no relation between missingness and outcomes; the sixth row presents results in which we allow the missing observations to be 3.5 percentage points less supportive of Obama than the observed respondents. The seventh row presents results from an inverse proportional weighting model that weights observed outcomes in a manner inversely proportional to their probability of being observed (Glynn et al. 2010). The eighth row presents results from Heckman’s well-know selection model (Heckman 1976).

Across all the models, the pro-Obama canvass appears to have decreased support for Obama by between -1.88 and -1.55 percentage points. These findings hold using methods that explicitly model selection (such as the Heckman selection model) and methods that impute or weight the data based on observed covariates. In

Table 5 Experimental conditions

		Canvass	No canvass
Mail	Phone	7,000	7,000
	No phone	7,000	7,000
No mail	Phone	7,000	7,000
	No phone	7,000	7,000

Number of households assigned to each experimental condition

this case, observed aspects of the selection process do not appear to be highly correlated with candidate preferences. As a result, various methods for dealing with sample selection—all of which make use of correlations among observed covariates in one way or another—converge on similar estimates.

Substantively, even the upper bounds for some of the most credible approaches are negative, and they are never larger than one-half of a percentage point. We can thus rule out all but the smallest positive effects of canvassing among this sample. What’s more, the negative effects of canvassing on Obama support are strongest among low-turnout voters, a group that is less engaged with politics and less easily mobilized by canvassing (see also Arceneaux et al. 2009; Enos et al. 2014). Being asked to vote for a specific candidate appears to be an unpleasant experience for a sizeable subset of our voters, one that makes them demonstrably less likely to respond to a separate survey and that appears to push them away from the sponsoring candidate. Whether that backlash is the product of the intensive campaign environment, a target universe with a disproportionate number of voters who live alone, or other contextual factors is a question for future research.

Conclusion

To ask someone to vote is to tap into widely shared social norms about the importance of voting in a democracy. To ask someone to vote for a particular candidate is a different story. In the words of a Wisconsin Democratic party chair, in persuasion, “[y]ou’re going to people who are undecided, who don’t want to hear from you, and are often sick of politics” (Issenberg 2012).

The results from the 2008 Wisconsin persuasion experiment illustrate just how difficult persuasion can be. Low-turnout voters appear to be turned off by in-person persuasion efforts. A single visit from a pro-Obama canvasser led some people to not respond to subsequent phone surveys, and appears to have pushed some to be less supportive of Obama. The relationship between a visit from a canvasser and survey response is strong and statistically robust. The relationship between the canvass treatment and reduced support for Obama is less statistically robust and requires stronger assumptions, but the results are still sufficiently precise that we can rule out all but minor positive effects. Backlash is more likely than persuasion.

The estimated persuasion effects are consistent across statistical methodologies. This suggests that the conditions for bias in estimating candidate support were not strongly satisfied, possibly because there was no common omitted variable that strongly influenced both the propensity to respond to the phone survey and the

Table 6 Omnibus balance tests

	Canvassed Full sample	Canvassed Survey respondents
Intercept	0.496* (0.021)	0.398* (0.046)
Obama expected support score	-0.018 (0.013)	-0.020 (0.027)
Turnout Score = 1	-0.003 (0.008)	0.042* (0.019)
Turnout Score = 2	0.003 (0.009)	0.046* (0.020)
Turnout Score = 3	0.013 (0.009)	0.082* (0.021)
Turnout Score = 4	0.002 (0.010)	0.087* (0.021)
Turnout Score = 5	0.009 (0.010)	0.081* (0.021)
Turnout Score = 6	0.001 (0.012)	0.072* (0.026)
Turnout Score = 7	0.003 (0.013)	0.052* (0.026)
Turnout Score = 8	-0.004 (0.015)	0.062* (0.030)
Turnout Score = 9	-0.022 (0.017)	0.035 (0.032)
Male	0.005 (0.004)	0.003 (0.009)
District % Dem.	0.003 (0.024)	0.041 (0.052)
Black	0.044* (0.016)	0.022 (0.036)
Hispanic	-0.010 (0.011)	0.010 (0.026)
Protestant	0.007 (0.005)	0.018 (0.010)
Catholic	0.013 (0.007)	0.025 (0.014)
District Med. household income	-0.000 (0.000)	0.000 (0.000)
District % college grads	0.008 (0.025)	0.006 (0.052)
<i>N</i>	55,980	12,439

Table 6 continued

	Canvassed Full sample	Canvassed Survey respondents
<i>F</i>	1.28	2.04

This table illustrates regressions of the canvassing treatment on 18 covariates to assess balance overall for both the full sample (left column) and the respondents who answered the survey (right column)

* $p < 0.05$

Table 7 Balance in random assignment

	Mean		<i>p</i> value	<i>N</i>
	Canvass assigned	Canvass not assigned		
Age	54.646	54.689	0.802	39,187
Black	0.021	0.018	0.037	56,000
Male	0.408	0.403	0.238	56,000
Hispanic	0.054	0.056	0.355	56,000
Voted 2002 general	0.206	0.204	0.523	56,000
Voted 2004 primary	0.329	0.329	0.943	56,000
Voted 2004 general	0.830	0.831	0.910	56,000
Voted 2006 primary	0.154	0.160	0.052	56,000
Voted 2006 general	0.551	0.550	0.786	56,000
Voted 2008 primary	0.356	0.351	0.254	56,000
Turnout score	2.865	2.862	0.861	56,000
Obama expected support score	47.629	47.893	0.102	55,990
Catholic	0.189	0.187	0.581	56,000
Protestant	0.453	0.450	0.405	56,000
District Dem. 2004	55.188	55.220	0.745	55,990
District Dem. performance - NCEC	58.476	58.528	0.571	55,990
District median income	45.588	45.524	0.558	55,980
District % single parent	8.563	8.561	0.948	55,980
District % poverty	6.656	6.690	0.558	55,980
District % college grads	19.282	19.224	0.534	55,980
District % homeowners	70.069	70.155	0.577	55,980
District % urban	96.712	96.843	0.161	55,980
District % white collar	36.074	36.040	0.638	55,980
District % unemployed	2.712	2.726	0.500	55,980
District % Hispanic	3.101	3.088	0.795	55,980
District % Asian American	0.809	0.823	0.288	55,980
District % Black	2.022	1.997	0.592	55,980
District % 65 and older	22.547	22.528	0.791	55,980

This table uses *t* tests to report the balance between those assigned to the canvassing treatment and those not assigned to the canvassing treatment for the full sample of respondents

Table 8 Balance in survey response assignment

	Phone treatment			Mail treatment		
	Mean		<i>p</i> value	Mean		<i>p</i> value
	Phone assigned	Phone not assigned		Mail assigned	Mail not assigned	
Age	55.706	55.924	0.519	55.577	56.051	0.161
Black	0.017	0.017	0.765	0.017	0.017	0.905
Male	0.394	0.391	0.672	0.395	0.390	0.536
Hispanic	0.041	0.046	0.200	0.045	0.042	0.448
Voted 2002 General	0.241	0.233	0.289	0.234	0.240	0.426
Voted 2004 Primary	0.389	0.373	0.068	0.378	0.383	0.579
Voted 2004 General	0.854	0.851	0.607	0.855	0.851	0.521
Voted 2006 Primary	0.194	0.186	0.278	0.194	0.185	0.209
Voted 2006 General	0.620	0.613	0.416	0.618	0.615	0.780
Voted 2008 Primary	0.426	0.409	0.043	0.419	0.416	0.753
Turnout score	3.245	3.168	0.062	3.203	3.210	0.863
Obama expected support	47.745	47.566	0.615	47.711	47.600	0.755
Catholic	0.182	0.178	0.637	0.179	0.181	0.711
Protestant	0.457	0.465	0.353	0.458	0.464	0.479
District Dem. 2004	54.754	54.767	0.949	54.742	54.779	0.860
District Dem.	58.094	58.098	0.984	58.069	58.124	0.779
District median income	46.180	46.019	0.480	46.109	46.090	0.933
District % single parent	8.229	8.241	0.873	8.198	8.273	0.337
District % poverty	6.308	6.315	0.953	6.286	6.336	0.680
District % college grads	19.591	19.776	0.350	19.742	19.625	0.556
District % homeowners	71.146	71.029	0.719	71.057	71.118	0.850
District % urban	96.783	96.815	0.868	96.951	96.647	0.116
District % white collar	36.413	36.183	0.135	36.297	36.299	0.987
District % unemployed	2.623	2.634	0.801	2.585	2.673	0.045
District % Hispanic	2.787	2.780	0.943	2.768	2.799	0.751
District % Asian American	0.803	0.787	0.573	0.784	0.806	0.436
District % Black	1.856	1.871	0.882	1.881	1.845	0.706
District % 65 and older	22.835	22.785	0.735	22.828	22.792	0.811

This table uses *t* tests to report the balance between those assigned to the phone and mail treatments and those not assigned to those treatments for individuals who answered the post-treatment phone survey in full

Table 9 Survey response rate differences across phone call treatment for all turnout levels

	N	Survey response rates			
		Phone call	No phone call	Difference	p value
0	5630	0.184	0.194	-0.010	0.352
1	13363	0.179	0.182	-0.004	0.569
2	10540	0.204	0.209	-0.005	0.513
3	7754	0.227	0.249	-0.023	0.018
4	6264	0.258	0.237	0.021	0.055
5	5273	0.273	0.259	0.014	0.267
6	2507	0.267	0.240	0.026	0.127
7	2210	0.274	0.294	-0.020	0.287
8	1406	0.319	0.253	0.066	0.006
9	1053	0.310	0.311	-0.002	0.949

This table reports the effect of being assigned to the phone call treatment on the probability of answering the post-treatment survey for each level of prior turnout, where zero indicates someone who has voted in no elections since 2000 and nine indicates someone who has voted in every primary and general election since 2000. The *p* values are estimated using *t* tests for each sub-group

propensity to support Obama. The contrast to the turnout analysis is noteworthy: in that case, civic mindedness or some correlate likely affected both turnout proclivity and responsiveness to the phone survey. As a result, we saw clear evidence of selection bias when analyzing the effects of canvassing on turnout. The magnitude of estimated backlash effect is approximately one to two percentage points. Note, however, that the experiment yields an intent-to-treat estimate. A standard result in the ITT literature is that the average treatment effect on treated individuals is the ITT divided by the contact rate, which in this case was 20 %. Given our ITT estimates, our evidence is consistent with as high as a 10 percentage-point reduction in stated Obama support for individuals who actually spoke to the canvassers.

There are several features of the experiment and its context that might limit the extent to which the results generalize. The experiment took place in October of a presidential election in a swing state, meaning that the voters in the study were likely to have been the targets of many other persuasion efforts. Also, the persuasive messages in the experiment emphasized economics, a central point in the 2008 campaign generally. For those reasons, the experiment tests the impact of persuasive messages that were already likely to be familiar. Moreover, the targeted universe focused on middle partisans in single-voter households, a group which may differ in its responsiveness to inter-personal appeals.

Still, this pattern of findings means that we need to tread carefully when analyzing experiments that involve separate, post-treatment surveys. Prior scholarship has illustrated that patterns of survey non-response can mean that estimates of campaign effects do not generalize to the broader population of interest (Vavreck 2007; Michelson 2014). Here, we build on that research by showing that even among survey respondents, estimated treatment effects can be biased. When the

Table 10 Non-parametric selection model results

	Full Sample	Prior turnout <3
Canvass	-0.0188 (0.0106)	-0.0266 (0.0167)
Phone call	-0.0105 (0.0103)	-0.0178 (0.0167)
Mail	0.0033 (0.0102)	0.0065 (0.0167)
Obama expected support score	0.0013 (0.0003)	0.0008 (0.0006)
Male	-0.0180 (0.0108)	-0.0275 (0.0176)
Age	-0.0006 (0.0004)	-0.0004 (0.0007)
District % Dem.	0.0011 (0.0006)	-0.0001 (0.0010)
Black	-0.0071 (0.0485)	0.0937 (0.0623)
Hispanic	0.0229 (0.0355)	0.0222 (0.0473)
Protestant	0.0090 (0.0115)	0.0217 (0.0190)
Catholic	0.0044 (0.0160)	0.0258 (0.0268)
Median income	0.0000 (0.0000)	-0.0000 (0.0000)
District % college grads	-0.0001 (0.0006)	0.0004 (0.0010)
Turnout score	-0.0028 (0.0029)	-0.0054 (0.0126)
Propensity	-2.8124 (1.8259)	-2.6938 (3.8032)
Propensity squared	4.6373 (2.7437)	4.4392 (6.1152)
Intercept	0.8979 (0.3113)	0.9957 (0.5906)
<i>N</i>	9415	3538

This table reports the results from non-parametric selection models in which the conditional expected outcome for the observed data is an additive function of the covariates and a correction term that depends on the estimated probability of being observed

Standard errors in parentheses

dependent variable is turnout or related outcomes, the fact that canvassing discourages low-turnout voters from even answering the phone is likely to induce bias. The treatment will look like it increased turnout by more than it actually did, as the treatment group will disproportionately lose low-turnout citizens relative to the control group.

When the dependent variable is vote intention, the direction of bias is less clear, but distortion could occur if, for example, anti-Obama voters were also the voters who became less likely to answer the phone survey after being canvassed. The surveyed treatment group would then appear more persuaded than it really was. At the same time, these results underscore the value of experimental designs that are robust to non-random attrition, including pre-treatment blocking (Nickerson 2005b; Imai et al. 2008; Moore 2012). Future experiments might also consider randomizing at the individual and precinct levels simultaneously (e.g. Sinclair et al. 2012) to provide a measure of vote choice that is observed for all voters.

Indeed, this experiment provides some broader methodological guidance for dealing with non-response in post-experimental surveys. As in other areas of causal inference, the best way to address such biases is to adopt experimental designs that are robust to them from the outset (Rubin 2008). For example, embedding experiments within panel surveys gives researchers the ability to match respondents before administering any treatments, and it provides more information about any non-random attrition as well. Making additional efforts to contact some respondents also might afford researchers benchmark estimates of treatment effects with reduced bias (e.g. Gerber et al. 2012). That said, there are nonetheless steps researchers can take to identify biases in traditional experiments with only post-treatment surveys. First, it is imperative to test for treatment effects on survey response, including assessments of possible heterogeneity across theoretically relevant subgroups. Researchers typically recognize that treatments may have differential effects on subgroups within a population; they should extend this logic to acknowledge that treatment effects on survey non-response are also likely to vary across sub-groups and are a potential source of bias. The concern, of course, is that in searching for treatment effects on survey responsiveness across subgroups, one may eventually identify a spurious effect due to the multiple comparisons involved. Therefore, this process should focus on theoretically relevant factors and should be conducted with statistical tests that appropriately account for the multiple comparisons being conducted (e.g. Westfall and Young 1993). More importantly, core experimental findings (including those presented here) should be replicated using out-of-sample tests to gauge their robustness and generality.

In cases where the treatments are found to influence survey response, there is no magic bullet. Researchers will have to invoke assumptions not justified by the experimental design alone, many of them quite strong. In those cases, rather than banking on any particular assumption, researchers are well advised to employ a variety of methods using different assumptions (Rubin and Schenker 1991). Imputation and weighting-based approaches are statistically efficient with excellent statistical properties as long as there are no unobservables affecting both the probability of response and the outcome of interest. Heckman models and more contemporary selection models can account for unobservables that affect both the probability of response and the outcome of interest, but tend to be fragile and depend heavily on exclusion restrictions that are rarely credible. Both of these modeling approaches make use of covariates that are predictive at the selection or outcome stages. Hence, to the extent that the researcher can influence data availability, she should make sure to gather a variety of high-quality covariates

when differential survey responsiveness across treatment groups is a potential concern.

Acknowledgments This paper has benefitted from comments by David Broockman, Kevin Collins, Eitan Hersh, Seth Hill, Michael Kellermann, Gary King, Marc Meredith, David Nickerson, Maya Sen, and Elizabeth Stuart. For research assistance, the authors gratefully acknowledge Julia Christensen, Zoe Dobkin, Katherine Foley, Andrew Schilling, and Amelia Whitehead. David Dutwin, Alexander Horowitz, and John Ternovski provided helpful replies to various queries. Earlier versions of this manuscript were presented at the 30th Annual Summer Meeting of the Society for Political Methodology at the University of Virginia, July 18th, 2013 and at Vanderbilt University's Center for the Study of Democratic Institutions, October 18th, 2013.

Appendix

Persuasion Script

Good Afternoon—my name is [INSERT NAME], I'm with [ORGANIZATION NAME]. Today, we're talking to voters about important issues in our community. I'm not asking for money, and only need a minute of your time.

As you are thinking about the upcoming election, what issue is most important to you and your family? [LEAVE OPEN ENDED—DO NOT READ LIST]

If not sure, offer the following suggestions:

- Iraq War
- Economy/ Jobs
- Health Care
- Taxes
- Education
- Gas Prices/Energy
- Social Security
- Other Issue

Yeah, I agree that issue is really important and that our economy is hurting many families in Wisconsin. Do you know anyone who has lost a job or their health care coverage in this economy?

I understand that a lot of families are struggling to make ends meet these days.

When you think about how that's affecting your life, and the people running for president this year, have you decided between John McCain and Barack Obama, or, like a lot of voters, are you undecided? [IF UNDECIDED] Are you leaning toward either candidate right now?

- Strong Obama
- Lean Obama
- Undecided
- Lean McCain
- Strong McCain

[If strong McCain supporter, end with:] Ok, thanks for your time this evening. [If strong Obama supporter, end with:] Great, I support Obama as well, I know he will bring our country the change we need. Thanks for your time this evening.

[ONLY MOVE TO THIS SECTION WITH LEANING OR UNDECIDED VOTERS] With our economy in crisis, job and health care loses at an all-time high, our country is in need of a change. But as companies are laying off workers and sending our jobs overseas, John McCain says that our economy is “fundamentally strong”—he just doesn’t understand the problems our country faces. McCain voted against the minimum wage 19 times. His tax plan offers 200 billion dollars in tax cuts for oil companies and big corporations, but not a dime of tax relief for more than a hundred million middle-class families. During this time of families losing their homes, McCain voted against measures to discourage predatory lenders and John McCain has never supported working families in the Senate and there is no reason to believe he will as President.

On the other hand, Barack Obama will do more to strengthen our economy. Obama will cut taxes for the middle class and help working families achieve a decent standard of living. Obama’s tax cuts will put more money back in the pockets of working families. He’ll stand up to the banks and oil companies that have ripped off the American people and invest in alternative energy. Obama will control the rising cost of healthcare and reward companies that create jobs in the U.S.

After hearing that, how are you feeling about our presidential candidates? What are your thoughts on this?

Obama will reward companies that keep jobs in the U.S., and make sure tax breaks go to working families who need them. Barack Obama offers new ideas and a fresh approach to the challenges facing Wisconsin families. Instead of just talking about change, he has specific plans to finally fix health care and give tax breaks to middle-class families instead of companies that send jobs overseas. Obama will bring real change that will finally make a lasting improvement in the lives of all Wisconsin families.

Now that we’ve had a chance to talk, who do you think you’ll vote for in November? John McCain and Barack Obama, or, are you undecided? [IF UNDECIDED] Are you leaning toward either candidate at this point?

- Strong Obama
- Lean Obama
- Undecided
- Lean McCain
- Strong McCain

Thanks again for your time, [INSERT VOTER’S NAME], we appreciate your time and consideration.

Survey Questions

“Hi, I’m calling with [survey firm redacted] with a brief, one-minute, opinion survey. We are not selling anything and your responses will be completely confidential.

Now first, thinking about the election for President this November, will you be voting for Senator Barack Obama, the Democratic candidate, or Senator John McCain, the Republican candidate?

1. Obama: Thank you. [GO TO Q2]
2. McCain: Thank you. [GO TO Q2]
3. VOLUNTEER ONLY Undecided/Don’t Know/Other: Thank You. [GO TO Q1]
4. VOLUNTEER ONLY REFUSED TO ANSWER [GO TO Q1]

If the election were held today and you had to decide right now, toward which candidate would you lean?

1. Obama
2. McCain
3. VOLUNTEER ONLY Completely Undecided
4. VOLUNTEER ONLY REFUSED TO ANSWER

Finally, for demographic purposes only, in what year were you born?” [Collect four-digit year]

Additional Tables

A Formal Statement of Selection Bias

Here, we formalize the problem of sample selection. Doing so enables us to group estimators based on their underlying assumptions about how fully the observed covariates can account for the patterns of missing data.

The dependent variable of interest is Y_i^* , support for Barack Obama. It is a function of the treatment (denoted as X_{1i}) and a vector of covariates (denoted as X_{2i}) that may or may not be observed. The treatment is randomized and is therefore uncorrelated with X_{2i} and the error terms in both equations below assuming a sufficient sample size.

$$Y_i^* = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

We only observe the Y_i^* for those voters who respond to the survey, indicated by the indicator variable d_i .

$$Y_i = Y_i^* d_i$$

The variable indicating that Y_i^* is observable is a function of the same covariates which affect Y_i^* .

$$d_i^* = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \eta_i$$

$$d_i = 1 \text{ if } d_i^* > 0$$

We assume the ϵ and η terms are random variables uncorrelated with each other and any of the independent variables.²² (Particular β or γ coefficients may be zero for variables that affect only selection or the outcome.)

We can re-write the equation for the observed data as

$$Y_i = Y_i^* |_{d_i=1}$$

$$= \beta_0 + \beta_1 X_{1i} |_{d_i=1} + \beta_2 X_{2i} |_{d_i=1} + \epsilon_i |_{d_i=1}$$

The various statistical approaches for dealing with sample selection diverge regarding their assumptions about X_{2i} . One common approach is to assume that X_{2i} is fully specified and observed. In such cases, we can predict the missing values for which $d_i^* < 0$ using the observed data. Statisticians refer to this assumption as “missing at random” (Schafer 1997; King et al. 2001; Little et al. 2002). Under this assumption, we might then apply some form of multiple imputation, which leverages the observed covariances among the variables to impute potential values for the missing data. Given that X_{2i} is fully specified, multiple imputation can be employed to estimate missingness in an outcome variable, an independent variable, or both.

Other approaches to sample selection are unwilling to assume that X_{2i} is fully observed—in such cases, the data are instead assumed to have non-ignorable missingness. These approaches turn to other assumptions, typically concerning the process that generates the missing data. If X_{2i} is unobserved, $\beta_2 X_{2i}$ will become part of the error term in the Y_i equation and $\gamma_2 X_{2i}$ will become part of the error term in the d_i equation. While X_{1i} (the randomized treatment) and X_{2i} are uncorrelated in the whole population, they are not necessarily uncorrelated in the sampled population. To see this, note that

$$X_{1i} |_{d_i=1} = X_{1i} |_{\gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \eta_i > 0}$$

$$X_{2i} |_{d_i=1} = X_{2i} |_{\gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \eta_i > 0}$$

The turnout case provides an example of how this bias can manifest itself. Suppose that the unobserved variable (X_{2i}) is unmeasured civic-mindedness, and it has a positive effect on whether someone responds to a pollster (implying $\gamma_2 > 0$) as well as a positive effect on Obama support (implying $\beta_2 > 0$). This would mean that in the observed data, the treated respondents would be more civically minded on average. Naturally, this could induce bias, as the treated, observed respondents are disproportionately high in civic-mindedness compared to observed respondents in the control group. This can explain the spurious finding in the surveyed-only column of Table 3. We know from the full data set that the treatment had no overall effect on turnout, but in the sub-sample of those who answered the follow-up survey, the

²² We could add additional covariates that only affect this equation without affecting our discussion below. The existence of such variables is commonly necessary for empirical estimation of selection models, although it is not strictly required, as these models can be identified solely with parametric assumptions about error terms.

canvass treatment is spuriously associated with a statistically significant positive effect.

Assuming X_{2i} is unobserved, two conditions must be met for sample selection to cause bias in randomized persuasion experiments with subsequent surveys:

1. $\gamma_1 \neq 0$. This is necessary to induce a correlation between randomized treatment and some unobserved variable in the observed sample. This can be tested and, for our data, we found $\gamma_1 < 0$ for low-turnout types and $\gamma_1 > 0$ for middle-turnout types.
2. $\gamma_2 \neq 0$ and $\beta_2 \neq 0$. In other words, given our characterization of the data-generating process, the error terms in the two equations are correlated.

If X_{2i} is not fully observed, the errors in the selection and outcome equations may be correlated. Heckman (1976) models such correlated errors by assuming that the errors in the two equations are distributed as bivariate normal random variables. This allows us to derive the value of the error term in the outcome equation conditional on being observed. Non-parametric selection models such as Das et al. (2003) approximate the conditional value of the error term with a polynomial function of the covariates. In practice, this involves fitting a first-stage model that produces a propensity of being observed. Powers of this fitted propensity are then included in the outcome equation.

Additional Estimation Strategies

Approximate Bayesian Bootstrap

Since non-random attrition threatens to bias listwise deletion models, we consider another imputation model that accounts for this possibility. In particular, we use hot deck imputation, which can be useful under three conditions satisfied by this experiment: when the missingness of interest is present primarily in a single variable, when the data contain many variables that are not continuous (Cranmer et al. 2013), and when there are many available donor observations (Siddique and Belin 2008b). Here, we employ the particular variant of hot deck imputation outlined in Siddique and Belin (2008b): an Approximate Bayesian Bootstrap (ABB) (see also Rubin and Schenker 1986; Rubin and Schenker 1991; Demirtas et al. 2007; Siddique and Belin 2008a). That approach has the added advantage that it can relax the assumption of ignorability in a straightforward manner by incorporating an informative prior about the unobserved outcomes.²³ These analyses focus on the 45,875 respondents who had Catalist phone match scores, although the results are similar when instead analyzing the full data set of 56,000 respondents.

Specifically, each iteration of the ABB begins by drawing a sample from the fully observed “donor” observations, which in our example number 12,439. This step

²³ Throughout these analyses, we drop our measure of respondents’ age, which is the only independent variable with significant missingness.

allows the ABB to more accurately reflect variability from the imputation. One can draw the donor observations with equal probability in each iteration, which effectively assumes that the missingness is ignorable conditional on the observed covariates. But importantly, researchers can also take weighted draws from the donor pool, which is the equivalent of placing an informative prior on the missing outcome data (Siddique and Belin 2008b). This allows researchers to relax the ignorability assumption, and to build in additional information about the direction and size of any bias.

Irrespective of the prior, we then build a model of the outcome using the covariates for the respondents with no missing outcome data, being sure to weight the donor observations by the number of times they were drawn in each iteration of the bootstrap. The subsequent step is to predict \hat{Y} for all observations—both donor and donee—by applying that model to the covariates X . For each observation with a missing outcome—there are 33,025 in this example—we next need to draw a “donor” observation that provides an outcome. Following Siddique and Belin (2008b), we do so by estimating a distance metric for each observation i as follows: $D_i = (|\hat{y}_0 - \hat{y}_i| + \delta)^k$, where δ is a positive number which avoids distances of zero.²⁴ For each missing observation, an outcome is imputed from a donor chosen with a probability inversely proportional to the distance D_i . As k grows large, note that the algorithm chooses the most similar observation in the donor pool with high probability, while a k of zero is equivalent to drawing any observation with equal probability.²⁵

Unlike a single-shot hot deck imputation, this approach does account for imputation uncertainty—and here, we fit our standard logistic regression model to 5 separately imputed data sets and then combine the answers using the appropriate rules (Rubin and Schenker 1986; King et al. 2001). Yet there is an important potential limitation to this technique. While running the algorithm multiple times will address the uncertainty stemming from the imputation of missing observations, it will not address the uncertainty stemming from small donor pools—and the reweighting in the non-ignorable ABB has the potential to exacerbate this concern (Cranmer et al. 2013).²⁶

We first run the Approximate Bayesian Bootstrap assuming ignorability (which means the prior is zero) and setting $k = 3$. Table 11 shows that, as we reported in the manuscript, such a model estimates the average treatment effect of canvassing to be -1.65 percentage points, with a corresponding 95% confidence interval from -3.29 to -0.01 . That estimate is similar to those recovered using listwise deletion. We also report additional results after adding an informative prior which reduces the share of respondents who back Obama from 57.5% in the observed group to 54.0% in the unobserved group. We chose the magnitude of the decline—3.5 percentage points—to approximate the largest decline in survey response observed across any of

²⁴ Here, δ is set to 0.0001.

²⁵ Siddique and Belin (2008a) report that a value of $k = 3$ works well in their substantive application, while Siddique and Belin (2008b) recommend values between 1 and 2.

²⁶ Still, even in light of this potential to under-estimate variance, Demirtas et al. (2007) demonstrate that the small-sample properties of the original ABB are superior when compared to would-be corrections.

Table 11 Overview of all results

Missing data strategy	Lower bound	50th	Upper bound
ABB, phone score, prior = 0, $k = 3$	-3.29	-1.65	-0.01
ABB, phone score, prior = 0, $k = 2$	-3.57	-1.89	-0.21
ABB, phone score, prior = 0, $k = 1$	-2.90	-1.34	0.23
ABB, phone score, prior = -3.5, $k = 3$	-3.34	-1.73	-0.05
ABB, phone score, prior = -3.5, $k = 2$	-3.52	-1.77	-0.02
ABB, phone score, prior = -3.5, $k = 1$	-2.67	-1.30	0.07
ABB, phone score, prior = -5.5, $k = 3$	-3.43	-1.76	-0.08
ABB, phone score, prior = -5.5, $k = 2$	-3.45	-1.75	-0.05
ABB, phone score, prior = -5.5, $k = 1$	-2.83	-1.27	0.28
ABB, all observations, prior = 0, $k = 3$	-3.69	-1.93	-0.17
ABB, all observations, prior = 0, $k = 2$	-3.47	-1.79	-0.11
ABB, All observations, prior = 0, $k = 1$	-2.83	-1.33	0.17

This table reports the lower bounds and upper bounds for several Approximate Bayesian Bootstrap estimations. The lower and upper bounds are the 2.5th and 97.5th percentiles of the average treatment effect. The units are percentage points

Note: “Phone score” refers to the 44,875 experimental subjects for whom a pre-treatment phone match score was available via Catalist. The prior indicates the level by which Obama support was adjusted in among unobserved respondents. As k increases, the preference for matching similar observations in the ABB increases

the turnout groups. In other words, in light of the differential attrition identified above, 3.5 percentage points is a large but still plausible difference between the observed and unobserved populations conditional on observed covariates. Here, the estimated treatment effect becomes -1.73 percentage points, with a 95% confidence interval from -3.34 to -0.05 . This result is essentially unchanged from the result with no prior. The table then presents various combinations of the prior and the k parameter, with little difference across the specifications except that reducing k below two (which means we are reducing the penalty for matching less similar observations) appears to increase the uncertainty regarding the estimated treatment effect. We also report results using all observations with, again, similar results.

Inverse Propensity Weighting

Inverse propensity weighting (IPW) is an alternative approach to dealing with attrition that uses some of the same building blocks as multiple imputation: it leverages information in the relationships among observed covariates to reweight the observed data such that they approximate the full data set (Glynn et al. 2010).

Specifically, we first use logistic regression on the full sample²⁷ to estimate a model of survey response. We employ the same model specification as above, with the exception that we drop our measure of age because it has substantial missingness. From the model, we generate a predicted probability of survey

²⁷ IPW requires data that are fully observed with the exception of the missing outcome. We thus set aside 20 respondents who were missing data for covariates other than age or Obama support.

Table 12 Heckman selection model results

	Full sample		Prior turnout <3
	Baseline model	with additional covariates	
Outcome equation			
Canvass	-0.016 (0.009)	-0.015 (0.009)	-0.036 (0.013)
Phone	-0.008 (0.009)	-0.008 (0.009)	-0.009 (0.013)
Mail	0.000 (0.009)	0.000 (0.009)	0.003 (0.013)
Constant	0.531 (0.027)	0.426 (0.036)	0.503 (0.052)
ρ	0.095 (0.043)	0.081 (0.044)	0.096 (0.057)
Selection equation			
Canvass	0.005 (0.013)	0.006 (0.013)	-0.05 (0.018)
Phone	0.004 (0.013)	0.004 (0.013)	-0.016 (0.018)
Mail	-0.005 (0.013)	-0.005 (0.013)	0.002 (0.018)
Weak phone match	0.759 (0.044)	0.772 (0.044)	0.79 (0.055)
Medium phone match	0.878 (0.028)	0.884 (0.028)	0.977 (0.036)
Strong phone match	1.108 (0.021)	1.107 (0.021)	1.117 (0.028)
Constant	-1.605 (0.023)	-1.592 (0.042)	-1.678 (0.060)
N-observed	12,442	12,442	5,647
N-censored	38,300	38,300	20,999

Standard errors in parentheses. The additional controls are predicted Obama support, district Democratic performance, male, Black and Hispanic

response for each respondent, estimates which vary from 0.13 to 0.35. For the 12,439 fully observed respondents, we then calculate the average treatment effect of canvassing, weighted by the inverse predicted probability of responding to the survey. Doing so, the estimated treatment effect of canvassing is -1.79 percentage points, with a 95% confidence interval from -3.52 to -0.05 percentage points.

Heckman Selection

Heckman selection models assume that the errors in the selection equation and outcome equation are distributed bivariate normally. With this assumption, the expected value of the error in the outcome equation conditional on selection can be represented with an inverse Mills' ratio. There is considerable disagreement in the

literature about the appropriateness of this assumption. Some find it implausible, given that the key assumption is about the joint distribution of unobserved quantities. Others find the approach more plausible than assuming away the correlation of errors across selection and outcome equations as is done in other selection models.

Table 12 shows results from several specifications of a Heckman selection model. In the first column no additional controls are included. In the second column, the controls listed at the bottom of the table are included. In the third column, the sample is limited to those who voted in 2 or fewer previous elections in the dataset. The results are qualitatively similar to the non-parametric selection model. The significant (or nearly so) ρ parameter indicates that there is some modest correlation between errors in the two equations. A statistically significant ρ parameter indicates that the errors are correlated, a necessary, but not sufficient condition for selection bias. In this case, since the estimates are similar to methods that assume no correlation of errors, there does not appear to be selection bias.

References

- Adams, W. C., & Smith, D. J. (1980). Effects of telephone canvassing on turnout and preferences: A field experiment. *Public Opinion Quarterly*, *44*(3), 389–395.
- Albertson, B., & Busby, J. W. (2015). Hearts or minds? Identifying persuasive messages on climate change. *Research & Politics*.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, *91*, 444–455.
- Ansolabehere, S., & Hersh, E. (2011). Who really votes? In P. M. Sniderman & B. Highton (Eds.), *Facing the challenge of democracy: Explorations in the analysis of public opinion and political participation*. Princeton University Press.
- Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, *20*(4):437–459.
- Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. *The Annals of the American Academy of Political and Social Science*, *601*(1), 169–179.
- Arceneaux, K. (2007). I'm asking for your support: The effects of personally delivered campaign messages on voting decisions and opinion formation. *Quarterly Journal of Political Science*, *2*(1), 43–65.
- Arceneaux, K., & Kolodny, R. (2009). Educating the least informed: Group endorsements in a grassroots campaign. *American Journal of Political Science*, *53*(4), 755–770.
- Arceneaux, K., & Nickerson, D. W. (2009). Who is mobilized to vote? A re-analysis of 11 field experiments. *American Journal of Political Science*, *53*(1), 1–16.
- Bechtel, M. M., Hainmueller, J., Hangartner, D., & Helbling, M. (2014). Reality bites: The limits of framing effects for salient and contested policy issues. *Political Science Research and Methods* (forthcoming).
- Broockman, D. E., & Green, D. P. (2014). Do online advertisements increase political candidates' name recognition or favorability? evidence from randomized field experiments. *Political Behavior*, *36*, 263–289.
- Cardy, E. A. (2005). An experimental field study of the GOTV and persuasion effects of partisan direct mail and phone calls. *The Annals of the American Academy of Political and Social Science*, *601*(1), 28–40.
- Cranmer, S. J., & Gill, J. (2013). We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, *43*(2), 425–449.
- Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33–58.

- Demirtas, H., Arguelles, L. M., Chung, H., & Hedeker, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics & Data Analysis*, 51(8), 4064–4068.
- Enos, R. D., Fowler, A., & Vavreck, L. (2014). Increasing inequality: The effect of GOTV mobilization on the composition of the electorate. *The Journal of Politics*, 76(1), 273–288.
- Enos, R. D., & Hersh, E. D. (2015). Party activists as campaign advertisers: The ground campaign as a principal-agent problem. *American Political Science Review*, 109(02), 252–278.
- Gerber, A., Karlan, D., & Bergan, D. (2009). Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2), 35–52.
- Gerber, A., & Green, D. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94(3), 653–663.
- Gerber, A. S., Kessler, D. P., & Meredith, M. (2011). The persuasive effects of direct mail: A regression discontinuity based approach. *Journal of Politics*, 73(1), 140–155.
- Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York, NY: W.W. Norton and Company.
- Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., & Hill, S. J. (2013). Who wants to discuss vote choices with others? Polarization in preferences for deliberation. *Public Opinion Quarterly*, 77(2), 474–496.
- Gerber, A. S., Huber, G. A., & Washington, E. (2010). Party affiliation, partisanship, and political beliefs: A field experiment. *American Political Science Review*, 104(04), 720–744.
- Gerber, A. S., Gimpel, J. G., Green, D. P., & Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *American Political Science Review*, 105(01), 135–150.
- Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1), 36–56.
- Green, D. P., & Gerber, A. S. (2008). *Get out the vote: How to increase voter turnout*. Washington, DC: Brookings Institution Press.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.
- Hersh, E. D. (2015). *Hacking the electorate: How campaigns perceive voters*. New York, NY: Cambridge University Press.
- Hersh, E. D., & Schaffner, B. F. (2013). Targeted campaign appeals and the value of ambiguity *The Journal of Politics*, 75(02), 520–534.
- 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(3), 769–781.
- Huber, G. A., & Arceneaux, K. (2007). Identifying the persuasive effects of presidential advertising. *American Journal of Political Science*, 51(4), 957–977.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observation lists about causal inference. *Journal of the Royal Statistical Society: Series A*, 171(2), 481–502.
- Issenberg, S. (2012). Obama Does It Better. *Slate*.
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49–69.
- Ladd, J. M., & Lenz, G. S. (2009). Exploiting a rare communication shift to document the persuasive power of the news media. *American Journal of Political Science*, 53(2), 394–410.
- Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14), 1355–1360.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
- Matland, R. E., & Murray, G. R. (2013). An experimental test for backlash against social pressure techniques used to mobilize voters. *American Politics Research*, 41(3), 359–386.
- Michelson, M. R. (2014). Memory and voter mobilization. *Polity*, 46, 591–610.
- Moore, R. T. (2012). Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4), 460–479.
- Nicholson, S. P. (2012). Polarizing cues. *American Journal of Political Science*, 56(1), 52–66.

- Nickerson, D. W. (2005a). Partisan mobilization using volunteer phone banks and door hangers. *The Annals of the American Academy of Political and Social Science*, 601(1), 10–27.
- Nickerson, D. W. (2005b). Scalable protocols offer efficient design for field experiments. *Political Analysis*, 13, 233–252.
- Nickerson, D. W. (2008). Is voting contagious? Evidence from two field experiments. *American Political Science Review*, 102(1), 49.
- Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, 21(2), 194–199.
- Panagopoulos, C., & Green, D. P. (2008). Field experiments testing the impact of radio advertisements on electoral competition. *American Journal of Political Science*, 52(1), 156–168.
- Rogers, T., & Nickerson, D. (2013). Can inaccurate beliefs about incumbents be changed? And can reframing change votes? HKS Faculty Research Working Paper Series RWP13-018.
- Rogers, T., & Middleton, J. A. (2015). Are ballot initiative outcomes influenced by the campaigns of independent groups? A precinct-randomized field experiment showing that they are. *Political Behavior*, 37, 567–593.
- Rogers, T., & Aida, M. (2014). Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased. *American Politics Research*, 42(3), 503–528.
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2, 808–840.
- Rubin, D. B., & Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, 10(4), 585–598.
- Rubin, D., & Schenker, N. (1986). Multiple imputation for interval estimation for simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), 366–374.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Shi, Y. (2015). Cross-cutting messages and voter turnout: Evidence from a same-sex marriage amendment. *Political Communication*. (forthcoming).
- Siddique, J., & Belin, T. R. (2008a). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, 27(1), 83–102.
- Siddique, J., & Belin, T. R. (2008b). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics & Data Analysis*, 53(2), 405–415.
- Simon, H. A., & Stern, F. (1955). The effect of television upon voting behavior in Iowa in the 1952 presidential election. *American Political Science Review*, 49(2), 470–477.
- Sinclair, B. (2012). *The social citizen*. Chicago, IL: University of Chicago Press.
- Sinclair, B., McConnell, M., & Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4), 1055–1069.
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769.
- Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.
- Vavreck, L., et al. (2007). The exaggerated effects of advertising on turnout: The dangers of self-reports. *Quarterly Journal of Political Science*, 2(4), 325–343.
- Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York, NY: Wiley.
- Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York, NY: Cambridge University Press.