Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts*

Michael A. Bailey[†] Daniel J. Hopkins[‡] Todd Rogers[§]

May 26, 2014

Can randomized experiments at the individual level help assess the persuasive effects of campaign tactics? To answer that question, we analyze a field experiment conducted during the 2008 presidential election in which 56,000 registered voters in Wisconsin were assigned to persuasive canvassing, phone calls, and/or mail. We find that persuasive appeals by canvassers had two unintended consequences. First, they reduced responsiveness to a follow-up survey among infrequent voters, a substantively interesting behavioral response that has implications for the statistical analysis of persuasion experiments. Second, the persuasive appeals possibly reduced candidate support and certainly did not increase it. This counterintuitive finding is reinforced by multiple statistical methods and suggests that contact by a political campaign can engender a backlash.

*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 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.

†Colonel William J. Walsh Professor of American Government, Department of Government and McCourt School of Public Policy, Georgetown University, baileyma@georgetown.edu.

‡Associate Professor, Department of Government, Georgetown University, dh335@georgetown.edu.

§Assistant Professor of Public Policy, Center for Public Leadership, John F. Kennedy School of Government, Harvard University, Todd_Rogers@hks.harvard.edu.

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 and Green, 2000; Green and Gerber, 2008; Nickerson, 2008; Arceneaux and Nickerson, 2009; Nickerson and Rogers, 2010; Sinclair, McConnell and Green, 2012), 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, 2005*a*; Arceneaux, 2007; Gerber, Karlan and Bergan, 2009; Gerber et al., 2011; Broockman and Green, 2013; Rogers and Nickerson, 2013). In these experiments, citizens are randomly selected to receive a message—perhaps in person, on the phone, in the mail, or online—and then they are surveyed alongside a control group whose members receive no message.

This paper assesses one such 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. A follow-up telephone survey then sought to ask all subjects about their preferred candidate, successfully recording the preferences of 12,442 registered voters.

We find no evidence that the persuasive appeals had their intended effect. Instead, the persuasive appeals had two unintended effects. First, persuasive canvassing reduced survey response rates among people with a history of not voting. Second, voters who were canvassed were *less* likely to voice support for then-Senator Obama, on whose behalf the persuasive efforts were taking place. 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 are consistent across a variety of statistical approaches and differ from other studies of political persuasion, both experimental (e.g. Arceneaux, 2007; Rogers and Middleton, 2013) and quasi-experimental (e.g. Huber and Arceneaux, 2007).

This paper highlights an unexpected methodological challenge for persuasion experiments that rely on follow-up surveys. We show that persuasive treatments can induce selection effects that need to be addressed in any causal analysis. To illustrate the potential for bias we show that failure to account for treatment-induced selection leads to demonstrably incorrect results when analyzing turnout.

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 selection can bias results, 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.

In the latter sections of the paper, we use various statistical models to analyze whether the campaign efforts persuaded voters to support Obama given the non-random attrition. These methods vary in their underlying assumptions: some assume that responses to the follow-up survey are predictable from observed covariates while others do not. Regardless of the model chosen, we find that the pro-Obama canvass had a borderline-significant, *negative* impact on Obama support of one to two percentage points. As a consequence, we can rule out even small positive persuasive effects of canvassing with a high degree of confidence. We conclude by summarizing the results and discussing ways in which they may or may not be generalizable.

1 Persuasion Experiments in Context

Political scientists have learned an immense amount about campaigns via experiments (Green and Gerber, 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 the effect of 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 many creative ways to study persuasion, a field experiment in which voters are randomly assigned to a treatment and then subsequently interviewed regarding their vote intention is particularly attractive, offering the prospect of high internal validity coupled with a real-world political context.[1] Moreover, by better understanding

[1]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, 2013) or discontinuities in campaigns' targeting formulae (e.g. Gerber, Kessler and Meredith, 2011).

persuasion, political scientists have the potential to shed light on voter decision-making as well as the nature of contemporary political representation.

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 encouraged to vote, they are being encouraged to do something that is almost universally applauded, giving inter-personal Get-Out-the-Vote efforts the force of social norms (Nickerson, 2008; Sinclair, 2012; Sinclair, McConnell and Green, 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 those that conflict with their prior views or partisanship (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) finds that phone calls and canvassing increase candidate support, and Gerber, Kessler and Meredith (2011) and Rogers and Middleton (2013) 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 and Kolodny (2009) show that targeted Republicans who were told that a Democratic candidate shared their abortion views nonetheless became less supportive of that candidate. Nickerson (2005*a*) finds no evidence that persuasive phone calls influence candidate support in a Michigan gubernatorial race, and Broockman and Green (2013) 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 reduce candidate support (Matland and Murray, 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.

Persuasion experiments also differ from turnout experiments in 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, Karlan and Bergan 2009, Gerber, Huber and Washington 2010, and Gerber et al. 2011). By the standards of contemporary survey research, such response rates are high. Still, there is little doubt that who responds is non-random. Given the high levels of non-response in prior studies of persuasion, sample attrition looms large as a possible source of bias.[2]

2 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 about 56% of the three million votes cast.

The experiment was implemented in three phases between October 9, 2008 and October 23,

[2] Experimental studies also rely on self-reported vote choice, not the actual vote cast. This is less of a concern, as public opinion surveys typically provide accurate measures of vote choice (Hopkins, 2009).

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 Catalist had a mailing 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 target population is thus likely to be less socially integrated on average, a critical fact given that two of the treatments involve inter-personal contact.

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. 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.[3]

In the second phase, every household in the target population 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 other groups received different combinations of the three treatments. The persuasive script for the canvassing and phone calls

[3]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.

was the same; it is provided in the Appendix. It involved an initial 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.[4] 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 about the "fundamentals of our economy."

Table B.1 in the Appendix indicates the division of voters into the various experimental groups. By design, each treatment was orthogonal to all others. 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. 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 did not report the outcome of individual-level voter contacts, meaning that our analyses must be intent-to-treat. Put differently, we do not observe what took place during

[4] Specifically, voters were coded as "strong Obama," "lean Obama," "undecided," "lean McCain," and "strong McCain."

the implementation of the experiment, and so are constrained to analyses which consider all subjects in a given treatment group as if they were treated. Subjects who were not home or did not answer the phone are included in our analyses, as are those who indicated strong support for a candidate and so did not hear the persuasive script in person or by phone.

The randomization appears to have been successful. Table B.2 in the Appendix shows means across an array of variables for subjects who were assigned to receive or not receive the canvass treatment. Of the 28 t-tests, only one returns a significant difference: subjects who are likely to be black according to a vendor-provided model are 0.3 percentage points more common in the group assigned to canvassing. That imbalance is small and chance alone should produce imbalances of that size in some tests. Similar results for the phone and mail treatments show no significant differences across groups.

In phase three, all voters in the targeted population 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 the year of their birth, and 85% of responses matched those provided by the voter file.

3 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 the canvassing treatment on survey response. While key covariates were balanced across the treatment and control groups in the full sample of 56,000, several politically important variables prove to be unbalanced across treatment and control groups among the 12,442 respondents who responded to the follow-up phone survey.

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 be canvassed and those not. Those who were assigned to canvassing were 1.9 percentage points more likely to have voted in the 2004 general election ($p = 0.03$), 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.[5]

The relationship between being canvassed and subjects' decision to participate in the telephone survey appears related to their prior turnout history. In Figure 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

[5] Table B.3 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 the biases are limited to treatments that involve interpersonal contact.

Table 1: **Balance among survey respondents.** 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.

	Mean			
	Canvass assigned	Canvass not assigned	P-value	N
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

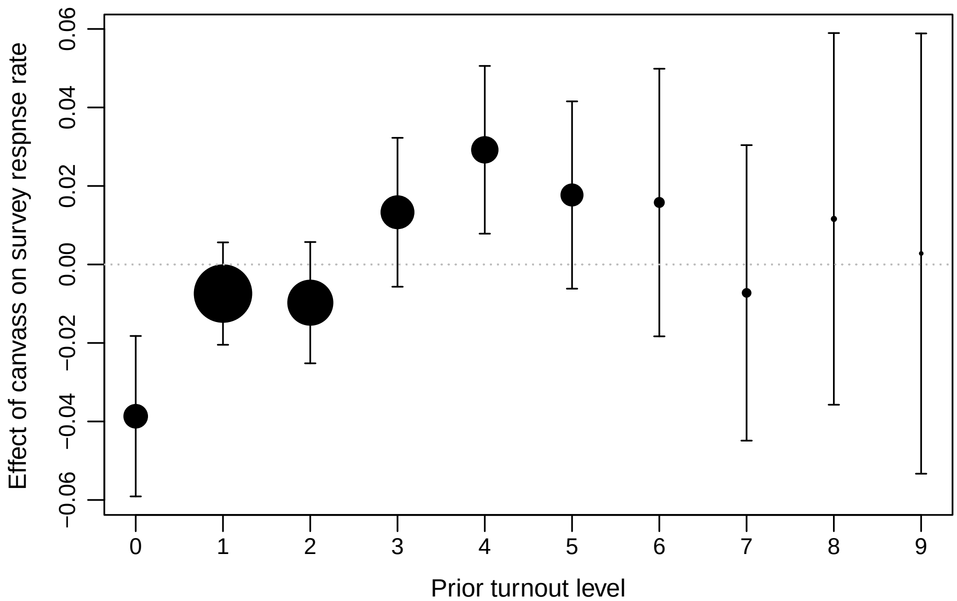

Figure 1: **Effect of canvass on survey response rates, by levels of prior turnout.** Each dot indicates the mean effect, and its size is proportional to the number of citizens in that group. The vertical lines depict the 95% confidence intervals.

prior election. The vertical lines span the 95% confidence intervals for each effect.[6]

Among the 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, with a p-value less than 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, canvassing has little discernible influence on survey response, although these groups account for few individuals in the experiment.[7]

[6] Voters under the age of 26 will 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.

[7] The effects for phone calls are generally similar, but not statistically significant (see Table B.4 in the Appendix). In results available upon request, we find no similar pattern of heterogeneous treatment effects on survey response

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 and yet increases the probability of survey response among those with middle levels of prior turnout. It is plausible that voters who infrequently vote 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 sometimes turnout are the most positively influenced by mobilization efforts (Arceneaux and Nickerson, 2009; Enos, Fowler and Vavreck, 2014), as ceiling effects limit the effect of mobilization among the most likely voters.[8]

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 makes clear,

for those who received campaign mailings.

[8] For example, Enos, Fowler and Vavreck (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.

the removal of those respondents leaves treatment and control groups that are well balanced in terms of their 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 same results hold true for the telephone treatment. The process of selecting households to call and calling them does not appear to have induced the biases identified above.

Table 2: **Breakdown of response differences.** This table reports the fraction of the previous nine elections in which respondents have voted, broken out by categories of response to the follow-up survey. The p-values are estimated using two-sided t-tests.

Sample	Mean Canvassed	Mean Control	Diff.	t-test p-value	N
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

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 pronounced bias, with the remaining members of the treated group having a higher prior turnout score than the control group by 0.008 (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. A similar pattern holds for receiving a persuasive phone call. There is no discernible bias in who answered the phone, but in the survey responses, those who were assigned to a persuasive phone call were 0.009 higher in the proportion of the nine previous elections in which they had voted. We found no such evidence for the mailing treatment. Whether in person or over the phone, persuasion attempts have a demonstrable effect on who participates in an ostensibly unconnected survey in the following weeks.

Selection Bias and Turnout

We have documented differential responsiveness to the survey—but does it 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 straightforward linear probability model to show that the canvass, phone, and mail treatments had no statistically significant effect on turnout in the 2008 November election for the full sample.[9] If we look only at those who responded to the survey, however, we get a different answer. The column on the right of Table 3 shows the result from the same model estimated only on those individuals who responded to the survey. Canvassing now appears to be associated with a 1.5 percentage-point increase in turnout. This ostensible effect is spurious and due entirely to selection. We know from the above discussion that the 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 sometimes but not always. This means that in the survey sample, we

[9] Results using logistic regression are highly similar.

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.

Table 3: **OLS Estimates of Effect of Treatments on Probability of Turnout.**

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

Standard errors in parentheses

The point is to demonstrate empirically that non-random attrition can matter. The experiment's sponsors did not intend, nor did we expect, the treatments to affect turnout. 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 the canvass treatment 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. These just weren't the behavioral response 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. Canvassing might even have decreased general election turnout among that group. 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.

4 Estimating Treatment Effects on Vote Intention

The goal of the persuasion campaign was to increase support for Barack Obama. To assess its effectiveness, we need to account for the non-random attrition detailed above. Here, we formalize the problem of sample selection and then briefly review the types of estimators most commonly applied to such problems. We group the 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. This 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^*.

$$
\begin{aligned}
d_i^* &= \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \eta_i \\
d_i &= 1 \text{ if } d_i^* > 0
\end{aligned}
$$

We assume the ϵ and η terms are random variables uncorrelated with each other and any of the independent variables. [10] (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

$$
\begin{aligned}
Y_i &= Y_i^*|_{d_i=1} \\
&= \beta_0 + \beta_1 X_{1i}|_{d_i=1} + \beta_2 X_{2i}|_{d_i=1} + \epsilon_i|_{d_i=1}
\end{aligned}
$$

The various statistical approaches for dealing with sample selection diverge as to the assumptions that they make with regard to 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 and Rubin, 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

[10] 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 assumption about error terms.

turn to other assumptions, typically about 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

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

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 observed treated respondents would be more civically minded on average. Naturally, this could induce bias, as the treated and observed respondents are disproportionately high in civic-mindedness when 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 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 across the two equations are distributed as bivariate normal random variables. This allows him to derive the value of the error term in the outcome equation conditional on being observed. Non-parametric selection models such as Das, Newey and Vella (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.

5 Results

In this section we analyze Obama support based on two statistical models that address non-random attrition in different ways. The first assumes data are missing at random, while the other allows for errors to be correlated across selection and outcome equations. We then provide an overview of extensive additional analyses using a variety of other statistical techniques to address sample selection.[11]

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

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 and Rubin, 2002), a technique which uses observed covariates to provide information about a respondent's likely response had she completed the survey. 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 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 (Samii, 2011, pg. 22).

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 fact affords it greater flexibility in its handling of variables that are not continuous, such as the binary outcome of interest here.[12]

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

[12] But that fact also means that the "implied joint distributions may not exist theoretically" (Buuren et al., 2006, pg. 1051). Still, that important theoretical limitation does not prevent MICE from working well in practice (Buuren et al., 2006).

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 in our model several other variables that could affect both whether and how an individual responded to the survey. We impute a Democratic support score to each respondent using Catalist's partisan support score, a continuous measure which draws on various demographic data and proprietary survey data. We control for gender, age, race, ethnicity and religion. The race and religion variables are from Catalist models predicting the likelihood a person is Black, Hispanic or Protestant. 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 fact 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. From the imputation model, researchers impute possible values of each missing observation, and then combine analyses of these data sets.[13]

As a baseline, we first estimate a model on 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

[13]To examine the performance of our model for multiple imputation using chained equations, we performed a series of five 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 25 resulting data sets was 74.4%, with a minimum of 71.4% and a maximum of 77.8%. This performance is certainly better than chance alone.

the analysis). The estimated difference in Obama support between those who were canvassed and those who were not was -1.6 percentage points (p=0.06, two-sided) controlling for the covariates listed above. This result suggests that if anything, canvassing made respondents *less* likely 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, which we do 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 -2.67, with a 95% confidence interval from -4.44 to -0.10 percentage points. Under this model, the persuasion effect of canvassing for the overall population was *negative*, and significantly so.[14] When we remove the 11,125 subjects who had no phone match score, we find that the treatment effect declines to -1.74.[15]

Given that canvassing had a negative effect on survey response (and potentially even turnout) among infrequent voters, it is valuable to examine its impact on support for Obama 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 -3.9 percentage points, with a 95% confidence interval from -7.3 to -2.2 percentage points. Here, we see stronger

[14] In fact, the associated p-value is less than 0.001, 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. To enable straightforward comparisons with the results from other models reported below, Table B.5 in the Appendix reports the results of a similarly specified linear probability model.

[15] The associated 95% confidence interval spans from -2.87 to 0.17.

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. For the other tactics, additional analyses (not shown here) find little evidence of persuasion in either direction. It appears as though a persuasive phone call or mailer does not produce the same backlash that an in-person visit does.

Non-Parametric Selection Model Next, we present results from the non-parametric, two-stage estimator for selection models detailed by Das, Newey and Vella (2003). The key difference from multiple imputation-type 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 B.6 in the Appendix displays the second-stage results for multiple specifications of the non-parametric selection model. For the full sample, the effect of canvassing is negative and borderline significant at -1.8 percentage points. The second column of Table B.6 shows that the effect is larger, but statistically insignificant, 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 are the results we have already discussed from the listwise deletion, multiple imputation, and non-parametric selection models. The additional rows summarize results we discuss further in the Appendix. The fifth and sixth rows present results from Approximate Bayesian Bootstraps (ABB) (Siddique and Belin, 2008*b*). 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. We discuss the ABB in greater detail and present additional results from it in the Appendix. 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 and Quinn, 2010; Samii, 2011). The eighth row presents results from Heckman's well-know selection model (Heckman, 1976). Finally, the ninth row presents Manski bounds which, due to the high rate of non-response, are not at all informative.

Table 4: **Overview of all results** This table reports the lower bounds and upper bounds for various estimators of the average treatment effect of canvassing. For the Manski bounds, the lower and upper bounds are sharp bounds. In all other cases, the lower and upper bounds are the 2.5th and 97.5th percentiles of the average treatment effect. The units are percentage points.

Missing Data Strategy	Lower Bound	50th	Upper Bound
Listwise Deletion – no covariates	-3.44	-1.63	0.09
MICE, all observations	-4.44	-2.67	-0.10
MICE, phone score	-2.87	-1.74	0.17
Non-Parametric Selection	-3.76	-1.80	0.21
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	-2.59	-1.78	-0.96
Heckman Selection	-3.29	-1.55	0.01
Manski Bounds, all observations	-78.14		77.42

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.

Across all the models (except the uninformative Manski bounds), the pro-Obama canvass appears to have decreased support for Obama by between -2.67 and -1.6 percentage points. These findings hold true using methods that explicitly model selection (such as the Heckman selection model) and methods that impute or weight the data based on observed covariates. This suggest that in this case, unobserved aspects of the selection processes are not highly correlated with candidate preferences.

Substantively, even the upper bounds for some of the most credible approaches are negative, and they are never larger than one-quarter 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 and Nickerson, 2009; Enos, Fowler and Vavreck, 2014). Being asked to vote for a specific candidate appears to be an unpleasant experience for at least 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.

6 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 appears to have led some people to not respond to subsequent phone surveys and to have pushed some people to be less supportive of Obama. Even persuasive phone calls appear to have influenced survey responsiveness.

The estimated persuasion effects are consistent across statistical methodologies. This implies that the conditions for bias in estimating candidate support were not strongly satisfied, likely because there was no common omitted variable that strongly influenced both the propensity to respond to the phone survey and the propensity to support Obama. The contrast to the turnout analysis is noteworthy: in that case, civic mindedness or a correlate likely affected both turnout proclivity and responsiveness to the phone survey. As a result, we saw clear evidence of selection bias there.

The magnitude of estimated backlash effects is approximately one to two percentage points. Note, however, that the experiment yields an intent-to-treat estimate. With a 20% contact rate, this implies that the actual canvassing effects could be as much as five times larger. In assessing that figure, readers should keep in mind that the length of the associated 95% confidence interval grows as well, and continues to include estimates quite close to zero. Moreover, if we use past research to develop a formal prior, it would almost certainly produce an estimate closer to zero than the mean average treatment effect on treated respondents implied by this experiment alone.

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. 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 of people who may have been less socially integrated and less responsive to inter-personal appeals than others.

Still, this pattern of findings means that we need to tread carefully when analyzing experiments that involve separate post-treatment surveys. When the dependent variable is turnout or related outcomes, the fact that the treatment 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 types 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 groups would then appear more persuaded than they really were. At the same time, these results underscore the value of experimental designs that are robust to non-random attrition, including pre-treatment blocking (Nickerson, 2005*b*; Imai, King and Stuart, 2008; Moore, 2012). Future experiments might also consider randomizing at the individual and precinct levels simultaneously (e.g. Sinclair, McConnell and Green, 2012) to provide a measure of vote choice that is observed for all voters.

References

Adams, William C. and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44(3):389–395.

Arceneaux, Kevin. 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, Kevin. 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, Kevin and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1):1–16.

Arceneaux, Kevin and Robin Kolodny. 2009. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53(4):755–770.

Broockman, David E. and Donald P. Green. 2013. "Do Online Advertisements Increase Political Candidates Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* Forthcoming.

Buuren, S. Van, J.P.L. Brand, C.G.M. Groothuis-Oudshoorn and Donald B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76(12):1049–1064.

Cardy, Emily Arthur. 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, Skyler J. and Jeff Gill. 2013. "We Have to Be Discrete about This: A Non-parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science* Forthcoming:1–25.

Das, Mitali, Whitney K. Newey and Francis Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *The Review of Economic Studies* 70(1):33–58.

Demirtas, Hakan, Lester M. Arguelles, Hwan Chung and Donald Hedeker. 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, Ryan D., Anthony Fowler and Lynn Vavreck. 2014. "Increasing Inequality: The Effect of

GOTV Mobilization on the Composition of the Electorate." *The Journal of Politics* 76(1):273–288.

Gerber, Alan, Dean Karlan and Daniel Bergan. 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, Alan and Donald Green. 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, Alan S., Daniel P. Kessler and Marc Meredith. 2011. "The Persuasive Effects of Direct Mail: A Regression Discontinuity Based Approach." *Journal of Politics* 73(1):140–155.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling and Seth J. Hill. 2013. "Who Wants to Discuss Vote Choices with Others? Polarization in Preferences for Deliberation." *Public Opinion Quarterly* 77(2):474–496.

Gerber, Alan S., Gregory A. Huber and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104(04):720–744.

Gerber, Alan S., James G. Gimpel, Donald P. Green and Daron R. Shaw. 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, Adam N. and Kevin M. Quinn. 2010. "An Introduction to the Augmented Inverse Propensity Weighted Estimator." *Political Analysis* 18(1):36–56.

Green, Donald P. and Alan S. Gerber. 2008. *Get Out the Vote: How to Increase Voter Turnout*. Washington, DC: Brookings Institution Press.

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selectionand Limited Dependent Variables, and Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5:475–492.

Hopkins, Daniel 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, Gregory A. and Kevin Arceneaux. 2007. "Identifying the Persuasive Effects of Presidential Advertising." *American Journal of Political Science* 51(4):957–977.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society: Series A* 171(2):481–502.

Issenberg, Sasha. 2012. "Obama Does It Better." Slate.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

Ladd, Jonathan M. and Gabriel S. Lenz. 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, Roderick J.A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, 2nd Edition*. New York, New York: John Wiley and Sons.

Manski, Charles F. 1990. "The Use of Intentions Data to Predict Behavior: A Best-Case Analysis." *Journal of the American Statistical Association* 85(412):934–940.

Matland, Richard E. and Gregg R. Murray. 2013. "An Experimental Test for Backlash Against Social Pressure Techniques Used to Mobilize Voters." *American Politics Research* 41(3):359–386.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science Experiments." *Political Analysis* 20(4):460–479.

Nicholson, Stephen P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1):52–66.

Nickerson, David W. 2005*a*. "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, David W. 2005*b*. "Scalable Protocols Offer Efficient Design for Field Experiements." *Political Analysis* 13:233–252.

Nickerson, David W. 2008. "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102(1):49.

Nickerson, David W. and Todd Rogers. 2010. "Do You Have a Voting Plan? Implementation Intentions, Voter Turnout, and Organic Plan Making." *Psychological Science* 21(2):194–199.

Panagopoulos, Costas and Donald P. Green. 2008. "Field Experiments Testing the Impact of Radio Advertisements on Electoral Competition." *American Journal of Political Science* 52(1):156–168.

Rogers, Todd and David Nickerson. 2013. "Can Inaccurate Beliefs About Incumbents be Changed? And Can Reframing Change Votes?" HKS Faculty Research Working Paper Series RWP13-018.

Rogers, Todd and Joel A. Middleton. 2013. "Are Ballot Initiative Outcomes Influenced by the Campaigns of Independent Groups? A Precinct-Randomized Field Experiment." HKS Faculty Research Working Paper Series RWP12-049.

Rubin, Donald B and Nathaniel Schenker. 1991. "Multiple Imputation in Health-care Databases: An Overview and Some Applications." *Statistics in Medicine* 10(4):585–598.

Rubin, Donald and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation for Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81(394):366–374.

Samii, Cyrus. 2011. "Weighting and Augmented Weighting for Causal Inference with Missing Data: New Directions." Working Paper, New York University.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Siddique, Juned and Thomas R. Belin. 2008*a*. "Multiple Imputation Using an Iterative Hot-deck with Distance-based Donor Selection." *Statistics in medicine* 27(1):83–102.

Siddique, Juned and Thomas R. Belin. 2008*b*. "Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data." *Computational Statistics & Data Analysis* 53(2):405–415.

Simon, Herbert A. and Frederick Stern. 1955. "The Effect of Television upon Voting Behavior in Iowa in the 1952 Presidential Election." *American Political Science Review* 49(2):470–477.

Sinclair, Betsy. 2012. *The Social Citizen*. Chicago, IL: University of Chicago Press.

Sinclair, Betsy, Margaret McConnell and Donald P Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments." *American Journal of Political Science* 56(4):1055–1069.

Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3):755–769.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York, NY: Cambridge University Press.

A 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 heath 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.

B Additional Tables

Table B.1: **Experimental conditions** Number of households assigned to each experimental condition.

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

Table B.2: **Balance in random assignment.** 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.

	Mean		p-value	N
	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	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

Table B.3: **Balance in survey response assignment** 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.

	Phone treatment Mean			Mail treatment Mean		
	Phone assigned	Phone not assigned	p-value	Mail assigned	Mail not assigned	p-value
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	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

Table B.4: **Survey response rate differences across phone call treatment for all turnout levels.** 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.

		Survey Response Rates			
	N	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

Table B.5: **Linear probability model results using multiple imputation.** Here, we fit linear probability models to the full data set after multiple imputation (left) as well as to the imputed data sets for all citizens who voted no more than 2 times in the prior 9 elections (right).

	Full sample	Prior turnout < 3
Canvass	-0.027	-0.045
	(0.011)	(0.019)
Phone call	-0.006	-0.007
	(0.012)	(0.018)
Mail	-0.006	-0.007
	(0.007)	(0.008)
Obama expected support score	0.0014	0.0013
	(0.0003)	(0.0004)
Male	-0.019	-0.020
	0.009	(0.012)
Age	-0.0012	-0.0012
	(0.0002)	(0.0002)
District Dem. performance	0.0012	0.0012
	(0.0008)	(0.0008)
Black	-0.019	-0.002
	(0.074)	(0.084)
Hispanic	-0.011	-0.006
	(0.035)	(0.039)
Protestant	0.017	0.016
	(0.011)	(0.011)
Catholic	0.018	0.020
	(0.008)	(0.012)
Median income	0.000	-0.000
	(0.000)	(0.000)
District % college grads	0.0004	0.0004
	(0.0004)	(0.0006)
Turnout score	0.0023	0.014
	(0.0016)	(0.008)
Constant	0.498	0.507
	(0.052)	(0.053)
N	56,000	29,533

Standard errors in parentheses.

Table B.6: **Non-parametric selection model results.** This table reports the results from a non-parametric selection model in which in 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.

	Full sample	Prior turnout < 3
Canvass	-0.018	-0.025
	(0.010)	(0.017)
Phone call	-0.010	-0.017
	(0.010)	(0.017)
Mail	0.003	0.007
	(0.010)	(0.017)
Obama expected support score	0.0013	0.001
	(0.0003)	(0.001)
Male	-0.015	-0.024
	(0.011)	(0.018)
Age	-0.001	-0.001
	(0.0003)	(0.001)
District Dem. performance	0.0011	-0.000
	(0.0006)	(0.001)
Black	-0.003	0.097
	(0.048)	(0.062)
Hispanic	0.017	0.020
	(0.035)	(0.047)
Protestant	0.007	0.021
	(0.011)	(0.019)
Catholic	0.003	0.025
	(0.016)	(0.027)
Median income	0.000	-0.000
	(0.000)	(0.000)
District % college grads	-0.000	0.000
	(0.001)	(0.001)
Turnout score	-0.003	-0.008
	(0.003)	(0.013)
Propensity	-1.542	-1.325
	(0.515)	(0.984)
Propensity squared	3.645	3.381
	(1.300)	(2.784)
Constant	0.651	0.723
	(0.077)	(0.122)
N	9,415	3,538

Standard errors in parentheses.

C 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 and Gill, 2013), and when there are many available donor observations (Siddique and Belin, 2008*b*). Here, we employ the particular variant of hot deck imputation outlined in Siddique and Belin (2008*b*): an Approximate Bayesian Bootstrap (ABB) (see also Rubin and Schenker, 1986, 1991; Demirtas et al., 2007; Siddique and Belin, 2008*a*). 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.[16] 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 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, 2008*b*). 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 (2008*b*), 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.[17] 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.[18]

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

[16] Throughout these analyses, we drop our measure of respondents' age, which is the only independent variable with significant missingness.

[17] Here, δ is set to 0.0001.

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

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 and Gill, 2013).[19]

We first run the Approximate Bayesian Bootstrap assuming ignorablility (which means the prior is zero) and setting $k = 3$. Table C.1 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. Adding add 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 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 and Quinn, 2010; Samii, 2011).

Specifically, we first use logistic regression on the full sample[20] 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 response for each respondent, estimates which vary from 0.13 to 0.36. 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 -1.78 percentage points, with a 95% confidence interval from -2.59 to -0.96 percentage points. Notice that IPW produces estimates with that are close to those using listwise deletion, and that have less variability then the estimates from MICE. This fact makes sense, as this version of the IPW approach does not include imputation uncertainty.

[19] 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.

[20] 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 C.1: **Overview of all results** 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.

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

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.

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 (Samii, 2011). 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 C.2 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 3 of 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.

Table C.2: **Heckman selection model results**

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

Standard errors in parentheses. Controls are included for predicted Obama support, district Democratic performance, male, Black and Hispanic.

Manski Bounds

As illustrated by Manski (1990), even in the case of missing outcomes, scholars can derive sharp upper and lower bounds for the average treatment effect. Specifically, we can make the most extreme possible assumptions about the missing outcomes and then estimate the potential average treatment effects under those assumptions. In one such scenario, we begin with the full data set of 56,000 voters. We then assume that everyone who was canvassed but not surveyed was behind McCain, while everyone who was not canvassed or surveyed backed Obama. If so, the estimated treatment effect is an extraordinary -78.14 percentage points. If we reverse the assumptions, such that canvassing induced every unobserved voter to support Obama and every uncanvassed voter supported McCain, the upper bound is 77.42 percentage points. When we are willing to make no assumptions beyond those inherent in the randomization in the presence of substantial attrition, we learn virtually nothing about the treatment effect.