Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts

Faculty Research Working Paper Series

Michael Bailey

Georgetown University

Daniel J. Hopkins

Georgetown University

Todd Rogers

Harvard Kennedy School

September 2013
RWP13-034

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

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

August 8, 2013

Can randomized experiments at the individual level help assess the persuasive effects of campaign tactics? In the contemporary U.S., vote choice is not observable, so one promising research design involves randomizing appeals and then using a survey to measure vote intentions. Here, we analyze one such field experiment conducted during the 2008 presidential election in which 56,000 registered voters were assigned to persuasion in person, by phone, and/or by mail. Persuasive appeals by canvassers had two unintended consequences. First, they reduced responsiveness to the follow-up survey, lowering the response rate sharply among infrequent voters. Second, various statistical methods to address the resulting biases converge on a counterintuitive conclusion: the persuasive canvassing reduced candidate support. Our results allow us to rule out even small effects in the intended direction and illustrate the backlash that attempts at inter-personal persuasion can engender.

*This paper has benefitted from comments by Kevin Collins, Seth Hill, Michael Kellermann, Gary King, Marc Meredith, David Nickerson, Maya Sen, and Elizabeth Stuart. For research assistance, the authors gratefully acknowledge Katherine Foley, Andrew Schilling, and Amelia Whitehead. David Dutwin, Alexander Horowitz, and John Ternovski provided helpful replies to various queries. An earlier version of this manuscript was presented at the 30th Annual Summer Meeting of the Society for Political Methodology at the University of Virginia, July 18th, 2013.

†Colonel William J. Walsh Professor of American Government, Department of Government and Public Policy Institute, 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. School of Government, Harvard University, Todd_Rogers@hks.harvard.edu.

Campaigns have two basic goals. They seek to mobilize and to persuade—to change who votes and to change 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; Gerber, Green and Larimer, 2008; Green and Gerber, 2008; Nickerson, 2008; Arceneaux and Nickerson, 2009; Nickerson and Rogers, 2010; Sinclair, McConnell and Green, 2012; Green, Aronow and McGrath, 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, or in the mail—and then they are surveyed alongside a control group whose members received no message. This paper assesses one such experiment, a 2008 effort in which 56,000 Wisconsin voters were randomly assigned to persuasive appeals on behalf of Barack Obama. Targeted registered voters were randomly assigned to persuasive canvassing, phone calls, and/or mailing. A follow-up telephone survey then sought to ask all subjects about

their preferred candidate, successfully recording the preferences of 12,442 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, and persuasion by phone appears to have done the same. Inter-personal persuasion can reduce responsiveness to a follow-up survey conducted by a different organization at a later time, suggesting that for some voters, it is an influential and negative experience. More unexpectedly, voters who were canvassed were *less* likely to voice support for then-Senator Obama, on whose behalf the persuasive efforts were taking place.

This paper highlights both methodological and substantive points about persuasion. Methodologically, the combination of random assignment to treatment and post-treatment surveys has been a primary tool for the assessment of individual-level persuasion. One of the core advantages of randomized experiments is that they permit causal inferences with assumptions that are limited and credible. Yet despite the promise and cost of this research design, it proves prone to bias in practice. Specifically, this paper demonstrates substantively meaningful selection effects induced by certain treatments. As a consequence, analyses of persuasion must be sensitive to the issue of sample selection, and must make assumptions that are significantly stronger than those justified by the randomization alone.

In the spirit of Rubin and Schenker (1991), we present the results from a variety of statistical techniques that address attrition through varying assumptions, from listwise deletion and multiple imputation (Schafer, 1997; King et al., 2001) to non-parametric bounding (Manski, 1990) and Approximate Bayesian Bootstraps (Rubin and Schenker, 1986, 1991; Siddique and Belin, 2008*b*,*a*).

The Appendix details additional results from parametric and non-parametric selection models (Heckman, 1976; Das, Newey and Vella, 2003) as well as inverse propensity weighting (Glynn and Quinn, 2010; Samii, 2011). Some of these approaches place added weight on observed cases that are similar to the unobserved cases, while others do the opposite. Yet despite invoking different assumptions to deal with non-random attrition, these methods are surprisingly consistent in their results. If anything, any outstanding biases are likely to push the effect further downward, since voters with lower baseline probabilities of supporting Obama might react more negatively to persuasive attempts on his behalf. As a consequence, we can rule out even small positive persuasive effects of canvassing with a high degree of confidence.

Substantively, these results illustrate how profoundly alienating politics can be for low-interest voters. A simple visit from a pro-Obama volunteer not only made these voters demonstrably less inclined to talk to a pollster on the phone, but also appears to have turned them away from Obama's candidacy. Door-to-door canvassing is known to be a powerful impetus to vote (e.g. Gerber and Green, 2000; Arceneaux and Nickerson, 2009)—but as these results make clear, persuasive canvassing can generate a backlash that reduces both survey participation and reported vote preference. The canvassing treatment examined here also left voter turnout unchanged.

These results are at odds with other studies of persuasion, both experimental (e.g. Arceneaux, 2007; Rogers and Middleton, 2013) and quasi-experimental (e.g. Huber and Arceneaux, 2007). As the results of field experiments on persuasion accumulate, scholars will be better positioned to assess the questions which this experiment raises but cannot answer, including how the volume of persuasive appeals influences persuasion. It is quite plausible that in mid-October of

2008, Wisconsin voters had already been saturated with attempts at persuasion, and so reacted negatively to yet another appeal. Also, the targeted voters in this experiment were exclusively registered voters living in households with no other registered voters, a group which is potentially less socially integrated and less responsive to inter-personal appeals than others.

In the next section, 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. Next, we present effects of the experiment on survey response, itself a meaningful and indicative pro-social behavior. In the subsequent section, we take into account non-random attrition and assess the efficacy of persuasion using the various statistical approaches before concluding.

Prior Research on Persuasion

How do American voters decide between presidential candidates, and can campaigns influence those decisions? These have been central questions in political science for decades (e.g. Lazarsfeld, Berelson and Gaudet, 1944; Campbell et al., 1960; Wlezien and Erikson, 2002; Johnston, Hagen and Jamieson, 2004; Brader, 2005; Lau and Redlawsk, 2006; Hillygus and Shields, 2008; Vavreck, 2009; Lenz, 2012). Contemporary social science has moved decisively toward testing theories with field experiments when possible, especially in the study of campaigns (e.g. Gerber and Green, 2000; Green and Gerber, 2003). Yet the vast majority of recent field experiments on campaign tactics have analyzed mobilization, not persuasion. The reason is simple: researchers cannot directly observe vote choice and therefore must "either randomly assign their treatments at the level of

the voting precinct, where vote choice can be counted in the aggregate, or gather individual-level data using a post-election survey"(Green and Gerber, 2008, pg. 159). Much of the recent evidence on the ability of campaigns to persuade has exploited natural experiments inherent in the uneven mapping of television markets to swing states (Simon and Stern, 1955; Johnston, Hagen and Jamieson, 2004; Huber and Arceneaux, 2007; Franz and Ridout, 2010) or the timing of campaign events (Johnston et al., 1992; Ladd and Lenz, 2009; Lenz, 2012). Other naturalistic studies have approached the problem through experiments with 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).

These methodological challenges notwithstanding, scholars cannot ignore persuasion. Campaigns for federal and statewide office expend substantial energy trying to find the most persuasive messages. They typically devote a majority of their resources to television advertising, a medium that is thought to persuade without mobilizing (Ashworth and Clinton, 2006; Huber and Arceneaux, 2007; Krasno and Green, 2008; Franz and Ridout, 2010). There is a long-standing literature within psychology on persuasion (e.g. O'Keefe, 2002), and extensive research on persuasion using survey and laboratory experiments (e.g. Brader, 2005; Chong and Druckman, 2007; Hillygus and Shields, 2008; Nicholson, 2012). Still, interest in experimental estimates of real-world persuasion has increased dramatically in recent years (Gerber, Karlan and Bergan, 2009; DellaVigna and Gentzkow, 2010; Barton, Castillo and Petrie, 2011; Gerber et al., 2011). By better understanding persuasion, political scientists have the potential to shed light on voter decision-making as well as the nature of contemporary political representation (Hill, 2010; Chong and Druckman, 2011;

Hersh, 2011).

Studying Persuasion versus Studying Turnout

Analyzing persuasion requires attention to the differences between studies of persuasion and turnout. Theoretically, there is almost universal agreement among Americans that turning out to vote is normatively good. Indeed, people who do not vote feel strong pressure to lie when asked about voting (Ansolabehere and Hersh, 2012). Voter turnout can thus be encouraged through the activation of social norms (Gerber, Green and Larimer, 2008; Nickerson, 2008; Sinclair, 2012; Sinclair, McConnell and Green, 2012). Face-to-face canvassing is among the most powerful techniques for increasing voter turnout, quite possibly because inter-personal interaction triggers those norms (e.g. Gerber and Green, 2000; Green and Gerber, 2008; Nickerson, 2008; Arceneaux and Nickerson, 2009).

In most cases, there is far less agreement on the question of *whom* one should support. What's more, in the contemporary U.S., the institution of the secret ballot reinforces the norm that voters need not disclose or discuss their choices. Gerber et al. (2013) report the widespread belief that vote choices are a personal matter, not one that should be discussed. That view is especially prevalent among voters who are less confident in their political abilities. Those who would seek to persuade do not have the advantage of drawing on widely shared norms in making their case. And when it comes to persuasion about vote choice, one of the advantages of personal appeals might be lost.

Those who would persuade voters face additional headwinds, as voters are liable to ignore or

reject appeals that are not consonant with their prior views or partisanship (Zaller, 1992; Taber and Lodge, 2006; Iyengar et al., 2008). In one survey experiment, Nicholson (2012) finds that campaign appeals do not influence in-partisans, but do induce a backlash among out-partisans. Similarly, in a field experiment, Arceneaux and Kolodny (2009) find that targeted Republicans who were told that a Democratic candidate shares their abortion views nonetheless became less supportive of that candidate. At the same time, Nickerson (2005*a*) finds no evidence that persuasive phone calls influenced candidate support in a Michigan gubernatorial race, and Broockman and Green (2013) find no evidence of persuasion through Facebook advertising. Null effects and even unintended effects are not uncommon, especially when the persuasive message cuts against voters' partisan predispositions.

Still, field experiments do detect persuasion in the intended direction under some circumstances. Studying the effects of television advertising, Gerber et al. (2011) find that the effects are demonstrable but short-lived. Both Gerber, Kessler and Meredith (2011) and Rogers and Middleton (2013) find that mailings influence support as intended by the sponsors. Yet in-person appeals seem not to have the same unique influence on vote choice that they do on voter turnout. For instance, in a study of a Democratic primary election for a New Mexico county commissioner in 2004, Arceneaux (2007) finds that both phone calls and canvassing had similar positive effects on candidate support. In short, experiments on persuasion produce conclusions that are frequently contingent.

Methodologically, a central difference between studies of turnout and studies of persuasion lies in how we observe the outcome. While turnout can be observed via administrative records,

individual-level persuasion studies are typically dependent on follow-up surveys. This dependence leads to two methodological challenges. The first is common to many studies of voting: we observe self-reported vote choice, not the actual vote cast. Still, public opinion surveys are typically effective at measuring vote choice, as their results match actual outcomes closely in most cases (Keeter et al., 2006; Hopkins, 2009).

Of greater concern is attrition (Samii, 2011; Gerber and Green, 2012; Little et al., 2012; Garciá, 2013). In individual-level persuasion experiments, we recover outcome data for only a subset of the experimental subjects, inducing concerns about sample attrition (Cardy, 2005). Adams and Smith (1980) report post-experimental survey response rates of approximately 63%, while for more recent studies, lower response rates including 25% (Arceneaux, 2007) and 32% (Gerber, Karlan and Bergan, 2009) are typical. Such response rates are high by the standards of contemporary survey research. But they have important implications for the subsequent analyses. Even if the attrition is random, the results only tell us about the sub-sample of respondents who completed the survey. Moreover, even small differences in attrition across experimental groups can induce substantial bias when the outcomes are not observed for upwards of two-thirds of all experimental subjects.

Wisconsin 2008

Here, we analyze a randomized field experiment undertaken by a liberal organization in Wisconsin in the 2008 presidential election. Wisconsin in that year was a battleground state, with approximately equal levels of advertising for Senators Obama and McCain.

The experiment was implemented in three phases between October 9, 2008 and October 23, 2008. In the first, the organization selected target voters who were "persuadable" Obama voters according to its vote model, lived in precincts that the organization could canvass, 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. 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 among these voters' likely partisanship, but with a clear concentration of so-called "middle partisans." 55% 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.[1]

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 treatments. The persuasive script for the canvassing and phone calls

[1]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.[2] 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 5 in the Appendix indicates the division of voters into the various experimental groups. By design, each treatment was orthogonal to the others. The organization implementing the experiment reported overall contact rates of 20% for the canvassing 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 are intent-to-treat. Put differently, we do not observe what took place during the implementation of the experiment,

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

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.

The randomization appears to have been successful. Table 6 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 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, 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.

Treatment Effects on Survey Response

Response patterns to the post-treatment survey are of substantial interest, both substantively and methodologically. Substantively, survey response is a pro-social behavior that yields insights into people's willingness to engage with strangers about politics (see especially Vigdor, 2004). Indeed,

as we will see in this section, the canvassing treatment appears to have similarly heterogeneous effects on survey response and on voter turnout. Methodologically, if our experimental treatments influence survey responsiveness, any naive analyses of the fully observed respondents are prone to bias. This section considers both issues by presenting balance tests and by analyzing the impact of the experimental treatments on survey response and voter turnout.

Table 1 shows balance tests for the subset of subjects who completed the telephone survey. There are marked, unexpected imbalances between canvass-assigned voters who answered the survey with regard to prior turnout. As compared to those not assigned to canvassing, 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 ($p = 0.01$). Since these imbalances do not appear in the full data set, this pattern suggests that canvassing influenced survey completion.

Table 7 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. It is noteworthy that the biases are limited to canvassing and phone calls, manipulations that involve interpersonal contact.

Subjects' decision to participate in the survey appears related to their prior turnout history.

Table 1: **Balance among survey respondents.** This table uses t-tests to report the balance between those assigned to canvassing treatment and those not for individuals who answered the post-treatment phone survey in full.

	Mean			
	Canvass assigned	Canvass not assigned	p-value	N
Age	55.756	55.875	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	1,2442
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	47.364	47.947	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	54.663	54.858	0.353	12,440
District Dem. performance - NCEC	58.010	58.183	0.374	12,440
District median income	46.262	45.937	0.155	12,439
District % single parent	8.186	8.284	0.212	12,439
District % poverty	6.219	6.404	0.127	12,439
District % college grads	19.791	19.576	0.279	12,439
District % homeowners	71.160	71.015	0.656	12,439
District % urban	96.640	96.959	0.099	12,439
District % white collar	36.309	36.287	0.882	12,439
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	22.817	22.803	0.921	12,439

To further understand the selection process at work, we divide our respondents into ten categories based on the number of prior elections since 2000 in which they had voted. Table 2 reports t-tests of the difference in survey response rates between canvassed and uncanvassed individuals in the post-treatment phone survey.[3]

While the response probabilities are generally increasing with prior turnout for both the treatment and control groups, the crucial difference is across experimental groups. Among the 5,630 respondents who have never previously voted, the canvassed individuals were a striking 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 similarly 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 four prior elections, it is sizable (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.

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 increases the probability of survey response among those with middle levels of prior turnout. As such, these heterogeneous treatment effects parallel the analyses of Enos, Fowler and Vavreck (2012), which show a non-monotonic relationship between turnout probabilities and responsiveness to mobilization.[4] The effects for phone calls are generally similar, but not statistically

[3]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. However, these young voters constitute only 5% of the low-turnout group.

[4]Specifically, Enos, Fowler and Vavreck (2012) estimates how the effects of direct mail, phone calls, and canvassing differ based on respondents' estimated probability of turning out to vote. It demonstrates that these campaign tactics have small effects for voters with low probabilities of voting, high effects for voters with middle-to-high

Table 2: **Survey response rate differences across canvass treatments for all turnout levels.** This table reports the effect of being assigned to canvassing 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 indicating someone who has voted in every primary and general election since 2000. The p-values are estimated using t-tests for each sub-group.

	N	Canvass	No canvass	Difference	P-value
0	5630	0.170	0.208	-0.039	0.000
1	13363	0.177	0.184	-0.007	0.265
2	10540	0.201	0.211	-0.010	0.217
3	7754	0.245	0.231	0.013	0.169
4	6264	0.262	0.233	0.029	0.007
5	5273	0.275	0.257	0.018	0.146
6	2507	0.262	0.246	0.016	0.364
7	2210	0.281	0.288	-0.007	0.706
8	1406	0.293	0.281	0.012	0.631
9	1053	0.312	0.309	0.003	0.923

significant (see Table 8 in the appendix). In results available upon request, we find no similar pattern of heterogeneous treatment effects on survey response for those who received campaign mailings.

While people who sometimes vote become more likely to take a survey after a campaign contact, people who seldom or never vote become less likely to do so. It is plausible that voters who infrequently vote find such interpersonal appeals bothersome, and so avoid the subsequent telephone survey. Darke and Ritchie (2007) report a related result on deceptive advertising, in which an initially deceptive advertisement induces distrust of advertising generally. At the same time, the persuasive contacts in our experiment trigger a pro-social response among those with middle levels of prior turnout. Such a response is consistent with prior research showing that those

probabilities of voting, and smaller but still positive effects for those with the highest probabilities of voting.

who sometimes turnout are the most positively influenced by mobilization efforts (Arceneaux and Nickerson, 2009; Enos, Fowler and Vavreck, 2012), as ceiling effects limit the effect of mobilization among the most likely voters.[5]

The differences in prior turnout by canvass treatment are not due to differences in the ease of contacting voters. Table 3 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, 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 call treatment. 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 3 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

[5]In explaining why treatment responsiveness might differ based on voters' probability of turning out, Enos, Fowler and Vavreck (2012, pg. 27) identifies differential contact rates as one explanation, as people with higher probabilities of voting are also easier to reach. In this case, however, the results cannot be explained entirely by differential contact rates among groups. Those who rarely or never voted in the past are not less responsive to treatment. Instead, they respond in the opposite way from those who sometimes vote.

Table 3: **Breakdown of response differences.** This table reports the fraction of the previous nine elections in which respondents have voted, broken out by categories of survey response. 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	0.362	0.351	0.011	0.016	12,399

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 subjects in the treated group who agree to take the survey are more likely to have turned out in prior elections than those in the control group. There are one or more correlates of prior turnout—a sense of civic duty, perhaps, or an interest in politics—that influence which treated respondents are willing to agree to the survey. Those with middle levels of prior turnout become more engaged and more likely to take the survey, while those with low levels of prior turnout are repelled. The bias doubles when examining the 12,399 respondents who actually reported a candidate preference, with the difference growing to 0.013 (p=0.005). Being canvassed leads higher-turnout respondents to be more likely to participate in the survey relative to the control. And part of this effect comes from respondents who are willing to take the survey but not to indicate a preference between presidential candidates.

A similar pattern holds for receiving a persuasive phone call, as Table 9 in the Appendix makes

clear. There is no discernible bias in who answered the phone, but in the survey responses, those who were called 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. Given that the treatments were explicit partisan appeals, it seems plausible that survey response might also have been influenced by partisanship, with Democrats more likely to have had a positive experience with the canvasser and thus more likely to respond to the subsequent survey. Yet we find little convincing evidence of that: the effect of the Democratic support score on survey response is positive but not significant in the data set as a whole. The same is true for the level of Democratic support in the precinct.

Treatment Effects on Voter Turnout

In other contexts, canvassing is known to be a powerful tool for encouraging voter turnout—and it can be analyzed straightforwardly, since turnout can be observed through post-election administrative records. Like survey response, voter turnout is a pro-social behavior, one which might give us another indication of whether inter-personal persuasion proved to be a negative experience. By estimating a simple logistic regression model of 2008 general election turnout on indicator variables for each of the treatment conditions, we get the another indication that this canvassing effort had distinctive effects from those detailed in prior experiments. For the population overall, none of the treatments is strongly predictive of turnout; for canvassing, the estimated effect is an 0.3 percentage point increase, with a 95% confidence interval from -0.4 to 1.1. Strikingly, canvassing is a near-significant *negative* predictor of turnout for those who have not voted in any of the prior 9 elections: the estimated effect is -1.3 percentage points, with a 95%

confidence interval from -2.9 to 0.4.[6] Whether these results differ from prior research because of the sample of single-voter households, the electoral context, or the nature of the persuasive appeal is unclear.

There are two important implications of the findings so far. First, the treatments did in fact induce behavioral responses. These just aren't the behavioral response expected. Those individuals who are least inclined to vote respond to a persuasive canvassing visit by becoming markedly less likely to complete a seemingly unconnected phone survey. In fact, 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 persuasion worked. In the next section, we discuss the challenges of heterogeneous treatment effects and differential survey response and present several approaches to estimating the efficacy of persuasion.

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 to account 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 treatment itself induced some low-turnout respondents to not respond while having the opposite effect among higher-turnout voters.

The general model underlying our statistical approaches to sample selection is standard. The

[6]The associated two-sided p-value is 0.12.

outcome, Y_i^* for every voter i is his or her support of Barack Obama. This is a function of the treatments (e.g. $Canvass_i$) and a vector of covariates X_i. We only observe the Y_i^* for those voters who respond to the survey, indicated by the dummy variable d_i.

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 Canvass_i + \beta_2 X_i + \epsilon_i \\ Y_i &= Y_i^* d_i \end{aligned}$$

The variable indicating we have observed voter i's view of Obama, d_i, is possibly a function of covariates which affect Y_i^* and covariates which only affect survey response Z_i:

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

The presence of missing data requires assumptions beyond those justified by randomization. And as Rubin and Schenker (1991) write in discussing non-random attrition, "[t]here is no *direct* evidence in the data to address the veracity of any such assumption, which is a good reason to consider several models and explore resultant sensitivity whenever possible"(Rubin and Schenker, 1991, pg. 588). Building on that advice, we employ multiple techniques that rely on differing assumptions to address sample selection. As a result, we are better positioned to evaluate the sensitivity of the estimated treatment effect to different assumptions about the process generating the missing data.

If $\gamma_1 = 0$, the error terms in the two equations are independent of each other, we can proceed

as if there is no missing data. In light of the discussion above, it is unlikely this is the case in this instance, although such analysis provides a baseline. If instead $\gamma_1 \neq 0$ and the error terms in the two equations are independent of each other, the missingness is random conditional on the independent variables. In the language of missing data, the unobserved data are "Missing at Random." In this case, we can use standard multiple imputation or inverse probability weighting approaches to recover the effect of the treatment.

If the error terms in the two equations are correlated, we have non-ignorable missingness. In this case, Approximate Bayesian Bootstrapping with a non-ignorable prior (Siddique and Belin, 2008*b*), the Heckman selection model (Heckman, 1976), and the nonparametric selection model (Das, Newey and Vella, 2003) make different assumptions to estimate treatment effects in samples subject to selection. We discuss the first technique below, and detail the latter two in the Appendix. These various approaches address missing data through different assumptions. But in different ways, they all get leverage from information in the observed covariates.

The potential impact of missing data is a function of how the outcome is measured as well as the number of observed and unobserved cases. In some models, we focus on subsets of the data set in which the level of missingness is lower. For example, Catalist provided a measure of the phone match quality for most respondents. There are 11,125 targeted voters for whom phone match scores were unavailable—and unsurprisingly, the survey response rate was dramatically lower among that group, at 5.3%. The phone match score was available prior to the treatment, and was in no way affected by it, meaning that removing respondents without scores introduces no bias.

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 who was 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, we learn virtually nothing about the treatment effect.

One way to tighten those bounds is by analyzing a subset of voters for whom response rates are higher. Analyzing only those 44,875 respondents with phone match scores, we can tighten the bounds marginally, to between -74.03 and 73.15. These bounds remain unhelpfully wide, ruling out only treatment effects which were already rendered implausible given the relatively low contact rates for canvassing. To provide substantively meaningful estimates, we will have to make additional assumptions.

Listwise Deletion

One alternative approach is to make the assumption that the data are Missing Completely at Random (MCAR)—that the missingness is unpredictable using either observed or unobserved measures. This assumption allows us to analyze the data as if they were fully observed. For example, the first column in Table 10 in the Appendix shows results from a logistic regression model using listwise deletion. The model includes only indicators for the three treatments: canvassing, phone, and mail. The model implies that a canvassed respondent is 1.63 percentage points less likely to indicate support for Obama, with a 95% confidence interval from -3.44 to 0.09. Substantively, it is highly surprising, as it suggests that the pro-Obama canvassing effort reduced support for Obama in the subsequent survey. For phone calls, the estimated treatment effect is -0.75, and the corresponding confidence interval runs from -2.61 to 0.89. So there, too, we see a negative point estimate, albeit with greater uncertainty. The effect of mailing is nearly zero, at -0.03 with a 95% confidence interval from -1.79 to 1.65. These would all be unbiased estimates were there no selection.[7]

Such a modeling approach can address selection, but only through its specification of covariates. In the second model in Table 10, we specify our fully saturated outcome model, which is a logistic regression model with 41 covariates that are pre-treatment and potentially related to survey responsiveness as well as Obama support. Based on the fact that survey responsiveness differs by prior turnout, we include 9 indicator variables for the number of prior elections in which the subject voted. We also include 3 indicator variables for each of the treatments, and another 18 indicator variables interacting the canvassing and phone call treatments with each of the prior

[7]We find no evidence of strong interactions among the treatments.

turnout indicators. This specification will address the specific selection problem described above, in which survey responsiveness varies by respondents' level of prior turnout. However, it seems likely that prior turnout is an imperfect proxy for the actual selection factor or factors, and thus that some selection biases will remain.

As additional covariates, our model also includes Catalist's partisan support score, a continuous measure which draws on various demographic data and proprietary survey data to impute a Democratic support score to each respondent. In addition, the model includes indicator variables for males and people who are likely to be black, Hispanic, Catholic, and Protestant according to Catalist models, and a continuous measure of age. 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. The second column in Table 10 illustrates select results from this model. Here, the estimated treatment effect is highly similar to that without covariates: -1.59 percentage points, with a somewhat wider 95% confidence interval from -3.68 to 0.51. Under the assumption of listwise deletion, the inclusion of covariates in a logistic regression does little to change the conclusion that canvassing likely had a *negative* average treatment effect. The effect also seems to vary with prior turnout. For respondents who have not turned out in any of the previous 9 primaries and general elections, the estimated treatment effect is -5.50, with an admittedly wide 95% confidence interval from -14.24 to 3.11 percentage points.

Multiple Imputation using Chained Equations

One common 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 makes use of observed covariates (such as a subject's partisan support score or attributes of her neighborhood) to provide information about her likely survey response had she completed the survey. Like many approaches to multiple imputation, our approach assumes 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 of the methods presented below, 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.[8] When employing multiple imputation, researchers typically develop a model or models of the relationship between each variable—

[8]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).

including pre-treatment and post-treatment measures—and every other variable. It is important to include any covariate which will be used in the estimation model in the imputation model as well. Our imputation model thus includes all of the variables described in the fully saturated model above.[9] It also includes the outcome measures. One is the outcome of primary interest, 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. We separately include a binary indicator of McCain supporters. From the imputation model, researchers impute possible values of each missing observation, and then combine analyses of these data sets.

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.

To estimate the treatment effects of persuasion, we then fit logistic regression models with the covariates detailed above to different data sets. For the 12,442 fully observed cases, the estimated difference in Obama support between those who were canvassed and those who were not was -1.6

[9]To simplify computation slightly, we include prior turnout as a single, continuous measure in both our imputation and outcome models in these analyses. Nonetheless, we continue to include interactions between prior turnout and the canvassing and phone call treatments.

percentage points (p=0.06, two-sided), suggesting that if anything, canvassing made respondents *less* likely to report supporting Obama. But given the results on survey response above, we might expect that that estimate is more of a lower bound. After all, it seems reasonable to suppose that those who do not support Obama were especially put off by the canvassing, and so differentially less likely 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. When we remove the 11,125 subjects who had no phone match score, we find that the treatment effect declines to -1.74.[10] Those respondents who are the hardest to reach are also potentially those who react more negatively to canvassing.

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 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 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

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

was undertaken. For the other tactics, analyses not shown 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.

Approximate Bayesian Bootstrap

The approach to multiple imputation using chained equations adopted above makes no distinction between dependent and independent variables at the imputation stage; it builds multivariate models for each variable with missingness in turn. Yet another approach to missing data—hot deck imputation—can be especially 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.[11] These analyses focus on the 45,875 respondents had Catalist phone match scores, although the results are quite 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

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

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 provide 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.[12] 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.[13]

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

[12]Here, δ is set to 0.0001.

[13]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).[14]

As a calibration exercise, we first run the Approximate Bayesian Bootstrap assuming ignorablility and setting $k = 3$. Across 50 iterations, we estimate the average treatment effect of canvassing to be -1.65 percentage points, with a corresponding 95% confidence interval from -3.29 to 0.01 (see the summary of all estimates in Table 4 below). That estimate is strikingly similar to those recovered using listwise deletion. Still, as Little et al. (2012) note, "[k]nowing the reasons for missing data can help formulate sensible assumptions about the observations that are missing"(1359). We then 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.

Assuming that the missing data are non-ignorable—and that the missing respondents are markedly less supportive of Obama than the observed respondents—we then re-estimated the ABB with k set to 3. Here, the estimated treatment effect becomes -1.73 percentage points,

[14]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.

with a 95% confidence interval from -3.34 to -0.05. This result is essentially unchanged from that above with no prior, and indicates that with k set to 3, the prior has little influence, as donor observations are still be matched to donees with highly similar predicted levels of Obama support.[15]

In Table 4, we present the results of various other ABBs, both ignorable (when the prior is set to zero) and non-ignorable. We also include ABBs estimated for the full data set of 56,000 respondents. The method produces confidence intervals that are consistently smaller than MICE, perhaps as a result of the variance issue identified above. In general, reducing k below 2 appears to reduce the estimated treatment effect, as lower values of k allow more dissimilar matches. With a data set of this size, and thus with a large number of available donor observations, the data appear to dominate the prior—at least for the values of the prior and k tested to date. But by comparing this technique to the Manski bounds, we gain insight into why that is the case. Whereas the Manski bounds make assumptions about the treatment effects, the prior we have implemented changes the composition of both the treatment and control groups. If every unobserved individual were anti-Obama, the estimated treatment effect would decline toward zero as the missing data grew large relative to the observed data.

[15]When we re-run the ABB setting k to 2, we are reducing the penalty for matching less similar observations. Yet when we do so while maintaining the same informative prior (a 3.5 percentage-point anti-Obama swing among the unobserved), we find a very similar result: -1.77 percentage points, with a 95% confidence interval from -3.52 to -0.02.

Discussion

Table 4 summarizes the results across the various methods for dealing with missing data employed here. To gain almost any information about the treatment's impact, we need to make assumptions beyond those justified by randomization alone: the width of the Manski bounds makes that clear. Still, the results prove surprisingly insensitive to the specific assumptions we make. The estimated treatment effect is generally stable across various methods commonly used to address missing outcome data, with the pro-Obama canvass seeming to have decreased support for Obama by between -2.67 and -1.6 percentage points.

As the Appendix details, the general finding holds true using still other methods, including those that explicitly model selection such as the Heckman selection model (Heckman, 1976) and the non-parametric selection model (Das, Newey and Vella, 2003). These estimators downweight those observed cases which are similar to the unobserved cases, while an estimator such as the IPW does the opposite. Yet the various approaches yield similar point estimates and levels of uncertainty. MICE does indicate a somewhat more negative treatment effect for the full data set, perhaps because of its capacity to impute missing age data.

What explains the surprising similarity between the estimates provided by methods that build upon differing assumptions? Several of these methods make use of covariates to model or condition on the process that leads some data to be missing. Such adjustments will only influence the average treatment effect to the extent that the covariates are related to selection and to treatment. Yet in this case, the analyses of survey response indicate that there was no strong interaction between respondents' partisanship and the treatments. More generally, the

Table 4: **Overview of 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
Manski Bounds, All Observations	-78.14		77.42
Listwise Deletion – No Covariates	-3.44	-1.63	0.09
Listwise Deletion – Covariates	-3.68	-1.59	0.51
MICE, All Observations	-4.44	-2.67	-0.10
MICE, Phone Score	-2.87	-1.74	0.17
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
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
Inverse Propensity Weighting	-2.59	-1.78	-0.96
Heckman Selection	-3.29	-1.55	0.01
Non-Parametric Selection	-3.4	-1.6	0.16

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.

similarity of the results across statistical approaches is consistent with the claim that whatever selection processes are at work are not highly correlated with candidate preferences. And the most plausible bias not accounted for by these methods is downward, if canvassed voters who are less supportive of Obama differentially avoided the subsequent survey.

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 and Nickerson, 2009; Enos, Fowler and Vavreck, 2012). Asking people to vote for a specific candidate is clearly an unpleasant experience for at least a sizable 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 valuable question for future research.

Conclusion

To ask someone to vote is to tap into widely shared social norms about the importance of voting and of 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). In keeping with that intuition, the results from the 2008 Wisconsin persuasion experiment illustrate just how difficult persuasion can be. Low-interest voters appear to be turned off of politics by in-person persuasion. They not only had lower probabilities of answering a follow-up telephone survey, but also exhibited lower support for Obama than similar voters who did not receive a canvassing visit.

What one thinks about the substantive import of these finding depends on what one was trying to get out of the experiment. If the goal is to explain voting, the treatment effect is clearly swamped by other factors. Even if one wants to explain survey response, the treatments examined here only change patterns of survey response at the margin. However, these results do shed light on what persuasion campaign is likely to do. Or at a minimum, they could suggest whether or not testing a program of persuasion is worthwhile. By that standard, we have learned a good deal. A single visit from a pro-Obama canvasser led some voters to not respond to subsequent phone surveys, an unexpected behavioral response that raises a methodological red flag when assessing the efficacy of the persuasion effort. These results become especially important as political campaigns incorporate randomized experiments into their targeting, as President Obama's campaign did in 2012 (Issenberg, 2012).

There are several features of the experiment and its context that are important to keep in mind when evaluating these results. 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 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 that is less politically engaged on average. And the outcome of interest is a survey response, making it conceivable that the actual effects on vote behavior differed from those observed in the survey.

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, 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 untreated 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 survey treatment groups in this instance would 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 such 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." *The Public Opinion Quarterly* 44(3):389–395.

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437–459.

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.

Ashworth, Scott and Joshua D Clinton. 2006. "Does Advertising Exposure Affect Turnout?" *Quarterly Journal of Political Science* 2(1):27–41.

Barton, Jared, Marco Castillo and Ragan Petrie. 2011. "What Persuades Voters? A Field Experiment on Political Campaigning.". Paper Presented at the 5th Annual Experimental Political Science Conference, New York University, March 2nd, 2012.

Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49(2):388–405.

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.

Campbell, Angus, Phillip E. Converse, Warren E. Miller and Donald E. Stokes. 1960. *The*

American Voter. New York: Wiley.

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.

Chong, Dennis and James N. Druckman. 2007. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101(04):637–655.

Chong, Dennis and James N. Druckman. 2011. Public-Elite Interactions: Puzzles in Search of Researchers. In *The Oxford Handbook of American Public Opinion and the Media*, ed. Robert Y. Shapiro and Lawrence Jacobs. New York, NY: Oxford University Press.

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.

Darke, Peter R. and Robin J.B. Ritchie. 2007. "The Defensive Consumer: Advertising, Deception, Defensive Process, and Distrust." *Journal of Marketing Research* XLIV(February):114–127.

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

DellaVigna, Stefano and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics* 2:643–669.

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. 2012. "Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate.". Mimeo, Harvard University.

Franz, Michael M. and Travis N. Ridout. 2010. "Political Advertising and Persuasion in the 2004 and 2008 Presidential Elections." *American Politics Research* 38(2):303–329.

Garciá, Fernando Martel. 2013. "Definition and Diagnosis of Problematic Attrition in Randomized Controlled Experiments." Manuscript, New York University.

Gerber, Alan, Dean Karlan and Daniel Bergan. 2009. "Does the Media Matter? A Field Ex-

periment 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. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton and Company.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Voter Turnout Experiment." *American Political Science Review* 102(1):33–48.

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., 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. 2003. "The Underprovision of Experiments in Political Science." *The Annals of the American Academy of Political and Social Science* 589(1):94–112.

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

Green, Donald P, Peter M Aronow and Mary C McGrath. 2012. "Field Experiments and the Study of Voter Turnout." *Journal of Elections, Public Opinion & Parties* 23(1):1–22.

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.

Hersh, Eitan. 2011. "Persuadable Voters in the Eyes of the Persuaders.". Mimeo, Yale University.

Hill, Seth J. 2010. "The Persuasion Region: A Theory of Electoral Change." Mimeo, University of California, San Diego.

Hillygus, D. Sunshine and Todd G. Shields. 2008. *The Persuadable Voter: Wedge Issues in Presidential Campaigns*. Princeton, NJ: Princeton University Press.

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 (statistics in society)* 171(2):481–502.

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

Iyengar, Shanto, Kyu S Hahn, Jon A Krosnick and John Walker. 2008. "Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership." *Journal of Politics* 70(1):186–200.

Johnston, R., A. Blais, H.E. Brady and J. Crête. 1992. *Letting the People Decide: Dynamics of a Canadian Election*. New York, NY: Cambridge Univ Press.

Johnston, Richard, Michael G. Hagen and Kathleen Hall Jamieson. 2004. *The 2000 Presidential Election and the Foundations of Party Politics*. New York, NY: Cambridge University Press.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70(5):759–779.

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.

Krasno, Jonathan S and Donald P Green. 2008. "Do televised presidential ads increase voter

turnout? Evidence from a natural experiment." *Journal of Politics* 70(1):245–61.

Ladd, Jonathan M.D. 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.

Lau, Richard R. and David P. Redlawsk. 2006. *How Voters Decide: Information Processing in Election Campaigns*. New York, NY: Cambridge University Press.

Lazarsfeld, Paul F., Bernard Berelson and Hazel Gaudet. 1944. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York, NY: Duell, Sloan, and Pearce.

Lenz, Gabriel S. 2012. *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. Chicago, IL: University of Chicago Press.

Little, Roderick J, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, James D Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung Shih, Jay Siegel and Hal Stern. 2012. "The Prevention and Treatment of Missing Data in Clinical Trials." *New England Journal of Medicine* 367(14):1355–1360.

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.

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.

O'Keefe, Daniel J. 2002. *Persuasion: Theory and Research*. Thousand Oaks, CA: Sage Publications.

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, John F. Kennedy School of Government, Harvard University.

URL: *http://scholar.harvard.edu/files/todd_rogers/files/are_ballot_initiative_outcomes_influenced_by_th*

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 withIgnorable 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." *The 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.

Vavreck, Lynn. 2009. *The Message Matters: The Economy and Presidential Campaigns*. Princeton, NJ: Princeton University Press.

Vigdor, Jacob L. 2004. "Community Composition and Collective Action: Analyzing Initial Mail Response to the 2000 Census." *Review of Economics and Statistics* 86(1):303–312.

Wlezien, Christopher and Robert S. Erikson. 2002. "The Timeline of Presidential Election Campaigns." *Journal of Politics* 64:969–993.

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 over seas, 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 5: **A. 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 6: **A. 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			
	Canvass assigned	Canvass not assigned	p-value	N
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
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 7: **A. 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. - NCEC	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
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 8: **Survey response rate differences across phone call treatment for all turnout levels.** This table reports the effect of being assigned to 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 9: **Breakdown of response differences for phone treatment.** This table reports the fraction of the previous nine elections in which respondents have voted, broken out by categories of survey response. The p-values are estimated using two-sided t-tests.

	Mean Turnout				
	Phone call	No phone call	Difference	p-value	N
Full Sample	0.318	0.319	-0.001	0.655	56,000
Record of Outcome	0.335	0.336	-0.001	0.759	41,808
+ Working Number	0.340	0.339	0.001	0.745	36,550
+ Participated in Survey	0.358	0.353	0.005	0.191	16,870
+ Reported Preference	0.361	0.352	0.009	0.047	12,399

Table 10: **Predicting Obama support via Logistic Regression**.

	(a)	(b)
Intercept	0.35***	0.19
	(0.04)	(0.27)
Canvass	-0.07*	-0.22
	(0.04)	(0.18)
Phone Call	-0.03	-0.26
	(0.04)	(0.18)
Mailing	0.00	0.01
	(0.04)	(0.04)
Canvass x 1 Prior Election		0.11
		(0.21)
Canvass x 2 Prior Elections		0.15
		(0.21)
Canvass x 3 Prior Elections		0.24
		(0.21)
Canvass x 4 Prior Elections		0.24
		(0.21)
Canvass x 5 Prior Elections		0.23
		(0.22)
Canvass x 6 Prior Elections		0.02
		(0.25)
Canvass x 7 Prior Elections		0.02
		(0.25)
Canvass x 8 Prior Elections		0.47*
		(0.28)
Canvass x 9 Prior Elections		0.14
		(0.29)
Num. obs.	12,442	9,415

Note: Standard errors in parentheses. * indicates significance at $p < 0.1$. The specification in column (b) also controls for being male, Black, Hispanic, Protestant, and Catholic as well as age and imputed partisanship. It further includes indicator variables for each level of prior turnout as well as interactions of the phone call treatment with each turnout category. At the aggregate level, the model includes the Census tract's median income and its percentage of college graduates. It also conditions on the precinct's Democratic performance in prior elections.

C. Alternative Estimators

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[16] 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.

[16]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.

Heckman Selection

Heckman selection models assume that the errors in the selection equation and outcome 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. This solution, while elegant, is implausible.[17] However, it may be no less implausible than assuming away correlation of errors. These models can provide another perspective on the treatment effects' sensitivity to particular assumptions.

In the first-stage Heckman model, we include two measures of the organization's coding of the quality of the phone number's match to the indicated individual. The presumption is that the better the quality of the phone match, the higher likelihood of reaching the individual. These variables are highly significant in the first stage of the model.

Table 11 presents results from two Heckman selection models. The results indicate that on average, canvassing had a weakly negative effect on views toward Obama, with an expected drop in probability of supporting Obama of 1.6 percentage points. Column (b) allows for heterogeneous treatment effects. The treatment is interacted with all of the possible values of prior turnout with the exception of turnout = 0, the excluded category. The coefficient on $Canvass$ is the effect of the treatment on voters who had not turned out in the 9 previous chances (spanning primary and general elections). This indicates that that the canvass treatment is associated with a decline of 3.8 percentage points in probability of supporting Obama. The effect for voters who turned out once in last nine voting opportunities was even more negative (5.3 percentage points, adding the

[17]For example, Samii (2011) notes that "[t]he rather extreme dependence on a model whose core feature—a model for the joint distribution of unobservable quantities—cannot be studied directly should raise some reasons for anxiety"(22).

Table 11: **Heckman selection model results**.

	(a)	(b)
Canvass	-0.016	-0.038
	(0.009)	(0.0256)
Mail	-0.0004	0.0001
	(0.009)	(0.0088)
Phone call	-0.008	-0.008
	(0.009)	(0.0089)
Predicted support		0.001
		(0.0003)
Male		-0.016
		(0.0093)
Democratic performance in district		0.001
		(0.0005)
Canvass x 1 Prior Vote		-0.015
		(0.0276)
Canvass x 2 Prior Votes		0.033
		(0.0285)
Canvass x 3 Prior Votes		0.031
		(0.0296)
Canvass x 4 Prior Votes		0.041
		(0.0311)
Canvass x 5 Prior Votes		0.054
		(0.0326)
Canvass x 6 Prior Votes		0.02
		(0.0393)
Canvass x 7 Prior Votes		0.009
		(0.041)
Canvass x 8 Prior Votes		0.059
		(0.0468)
Canvass x 9 Prior Votes		0.008
		(0.0517)
(Intercept)	0.53	0.398
	(0.27)	(0.0426)
ρ	0.095	0.088
	(0.04)	(0.0444)

Standard errors in parentheses. Additional controls for turnout score, race and ethnicity (Black and Hispanic indicators), religion (Catholic and Protestant indicators), and the percentage of college graduates in tract were included but are not reported here.

main effect and the interaction effect). For other voters, the interaction effect generally offsets the negative coefficient on the main effect.

Nonparametric Selection

Das, Newey and Vella (2003) provide a nonparametric approach to sample selection. In the first stage, it uses a series estimator of the selection probability, and in the second stage it conditions on various functions of the selection probability. In practice, this entails estimating a propensity score in the first stage and in the second stage including a polynomial function of the propensity score as a control. Selection of specific functional forms takes place via cross-validation based on minimization of forecast errors when all other observations are used to predict each single observation.

In the first stage of the nonparametric model, we use all the variables from Table 11. We also use three additional variables which are related to the vendor-assessed quality of the phone number information. We are assuming that these factors explain whether or not someone answered the phone survey but do not, conditional on the other variables in the model, explain vote intention.

Table 12 shows essential results from the second stage of the nonparametric selection model. In model (1), the treatments are not interacted with turnout history. Consistent with earlier results, the treatment effect is to reduce support for Obama by about 1.6 percentage points. In model (2), we interact the canvass treatment with all turnout categories from 1 to 9, meaning that the coefficient on canvass is the effect for people with no recorded prior turnout. Again, consistent with earlier results, the treatment appears to lower their support for Obama by 3.3

percentage points. The fact that the propensity scores are statistically insignificant suggests that selection is not causing bias. We have estimated models with up to fifth-order polynomials in propensity scores, with no substantial changes in results.

Table 12: **Nonparametric selection model**.

	Model 1	Model 2
Canvass	-0.016	-0.033
	(0.009)	(0.029)
Phone call	-0.007	-0.022
	(0.009)	(0.027)
Mail	-0.000	0.000
	(0.009)	(0.009)
Propensity score	0.156	0.012
	(0.144)	(0.173)
N	12,439	12,439
R^2	0.005	0.006

Standard errors in parentheses. * indicates significance at $p < 0.05$. Both models have additional control variables as in Table 11. Model 1 has no interactions with canvassing; model 2 interacts canvassing with prior turnout.